Langages classificatoires et recherche d'information sur les portails d'entreprise : quels apports pour les utilisateurs ? Ex : Les taxinomies du portail Intralignes d'Air France.

Gaelle Le Targat

sciences
et
techniques
de
l'information

INTD

CONSERVATOIRE NATIONAL DES ARTS ET METIERS
INSTITUT NATIONAL DES TECHNIQUES DE LA DOCUMENTATION

MEMOIRE

en vue obtenir le

DESS en Sciences de l'information et de la documentation spécialisées

Présenté et soutenu par

Gaëlle LE TARGAT

Le 13 Octobre 2005

Langages classificatoires et recherche d'information sur les portails d'entreprise : quels apports pour les utilisateurs ?

Ex : Les taxinomies du portail Intralignes d'Air France.

Jury
Bernard BONNET (ONISEP / INTD)
Hervé MARCHAND (Air France)

Cycle supérieur Promotion XXXV

Remerciements

Je remercie Hervé Marchand pour ses précieux conseils qui m'ont permis de découvrir les arcanes d'Air France, et plus largement pour son « kit » de survie au sein d'une grande entreprise.

Merci à Bernard Bonnet, pour avoir accepté de diriger cette étude et pour ses conseils avisés tout au long de sa réalisation.

Merci également à Audrey Blanchard pour son soutien et le temps qu'elle m'a accordé.

Je tiens aussi à remercier Pascal Matthieu, responsable du DP.SQ, Bernard Montagne et toute l'équipe du DP.SQ pour leur accueil chaleureux et leur coopération.

Pour finir, je souhaiterai remercier ma famille pour l'aide qu'ils m'ont apportée et la patience dont ils ont fait preuve tout au long de cette année.

Résumé

Après un rappel des notions fondamentales qui régissent la recherche d'information et les processus d'indexation. L'auteur évoque l'origine des langages documentaires, en s'attardant plus particulièrement sur les langages classificatoires. Il dresse ensuite une typologie de ces outils en étudiant l'évolution de leurs fonctions au contact des systèmes de gestion de contenus informatisés. Il s'attache plus particulièrement à définir les apports des moteurs de recherche linguistiques, pour la gestion des portails intranet.

Puis, dans la seconde partie de ce mémoire, l'auteur s'intéresse à l'émergence de nouveaux outils de recherche et d'indexation : les ontologies, les taxinomies, et les topic maps. Il étudie leurs fonctionnalités en les mettant en perspective avec celles des langages classificatoires traditionnels, et recense leurs différents contextes d'utilisation. Puis, il examine de quelle manière ces outils permettent d'améliorer la pertinence des recherches, et l'accès aux informations utiles pour le personnel des entreprises.

Il illustre ces apports par l'étude d'une taxinomie d'entreprise, conçue pour structurer et guider les recherches effectuées sur le portail Intralignes de la compagnie Air France.

Mots-clés :

TAXONOMIE, ONTOLOGIE, CARTE DE TOPIQUES, THESAURUS, TALN, INDEXATION AUTOMATIQUE, SOURCE D' INFORMATION, SYSTEME D'INFORMATION, DSI, CLASSIFICATION, LANGAGE DOCUMENTAIRE, RECHERCHE D'INFORMATION, PORTAIL, INTRANET, ENTREPRISE, MOTEUR DE RECHERCHE, CATEGORISATION, RECHERCHE EN LIGNE.

Table des matières

PARTIE 2 :

NOUVELLES FONCTIONNALITES DES LANGAGES CLASSIFICATOIRES : DE L'ACCES A L'INFORMATION A LA NAVIGATION INFORMATIONNELLE. 65

PARTIE 3 :

LE PORTAIL « INTRALIGNES » D'AIR FRANCE : UN EXEMPLE DE « CORPORATE TAXONOMY »

Liste des figures

Liste des annexes

Introduction

En investissant les entreprises puis les foyers individuels, l'Internet a mis à porter de tout un chacun une masse d'informations jusqu'alors difficilement consultables sans l'aide des documentalistes.

En quelques décennies, la machine à écrire a été remplacée par l'ordinateur, et le document électronique a supplanté le papier. Fini le mémo que l'on déposait sur votre bureau, aujourd'hui, vous voila informé via un courriel envoyé simultanément à l'ensemble des employés concernés !!! Ainsi, le document électronique, en apparence plus facile à créer qu'un document papier (mais aussi plus économique, puisqu'il peut être corrigé indéfiniment avant impression), a connu un développement incomparable au sein des entreprises.

Face au volume et à l'hétérogénéité des documents produits, les organisations ont tenté de rationaliser le stockage et la circulation de ces informations en mettant en place des bases de données, des réseaux intranet et enfin des portails pour organiser l'accès à ces sites. Plébiscités par les éditeurs de logiciels, ces outils sont très vite apparus, comme « la solution miracle » pour améliorer la gestion des documents produits au sein des organisations.

Toutefois, sur le web comme dans les entreprises, les utilisateurs ont rapidement été confrontés aux problèmes générés par cette surabondance de renseignements[1]. En effet, stocker des documents n'est pas une solution en soi, encore faut-il pouvoir les retrouver rapidement quand on en a besoin !

Isolé face à son moteur de recherche, souvent peu aux faits des principes de fonctionnement de cet outil et ignorant des règles de base concernant les langages d'interrogation et d'indexation, l'utilisateur « lambda » se retrouve fréquemment noyé sous une multitude de résultats ou a contrario dépité face à une réponse négative à sa requête alors qu'il sait que le document existe !!!

Conçus pour et par des documentalistes, la complexité des langages documentaires a incité les éditeurs de logiciel à développer des moteurs de recherche permettant une

[1] Selon une enquête menée par Altavista en 2001, le volume d'information généré par un salarié double tous les 18 mois.

interrogation en langage naturel[2], plus adaptés aux utilisateurs néophytes. Si la formulation des requêtes est devenue plus aisée, la pertinence des résultats et surtout leur tri restent encore à améliorer.

C'est pourquoi les concepteurs de moteurs de recherche, travaillent depuis quelques années déjà, à l'amélioration de leurs outils de « retrouvage » (retrieval system) en tentant de coller au plus près à la démarche cognitive des « non spécialistes de l'information » lorsqu'ils effectuent une recherche. [9 –MANIEZ, Note 1, p.15]
L'une des pistes explorées, s'oriente vers la personnalisation de l'accès aux informations recherchées. S'appuyer sur le profil des utilisateurs pour améliorer la pertinence des résultats obtenus et « l'orientation de l'usager vers la bonne ressource informationnelle », voila le nouveau défi posé aux éditeurs !! [73-DALBIN],
C'est dans ce contexte que les langages classificatoires semblent aujourd'hui revenir sur le devant de la scène. Révélatrice de l'intérêt, mais aussi du «flou» qui entoure ces notions dans l'esprit des professionnels, on constate depuis quelques années, une forte augmentation des articles consacrés aux taxinomies, ontologies et topics dans la presse anglo-saxonne spécialisée.

A travers, un projet de création de portail documentaire pour le service DP.SQ d'Air France, ce mémoire s'attachera à montrer l'intérêt d'une « réintégration » des langages classificatoires dans les systèmes de recherche actuels, comme outil de gestion de la pertinence.

Dans un premier temps, nous étudierons en quoi consiste la recherche d'information et quels sont les processus et notions qui se dissimulent derrière cette démarche.

Ainsi, nous verrons comment la nécessité de définir le contenu d'un document pour améliorer la pertinence des recherches, a nécessité la création des langages d'indexation.

Après un bref rappel historique sur l'origine et les fonctions des langages documentaires, nous établirons un inventaire de ces outils en nous attachant plus particulièrement aux langages classificatoires et à leur spécificité. Puis, nous

[2] Recherche full text indexant l'intégralité des mots d'un document , tout en incluant des couches de traitements linguistiques pour affiner les relations sémantiques entre les termes.

aborderons les outils et les processus de fonctionnement de la recherche documentaire informatisée (RDI). Nous examinerons plus particulièrement les portails, et les moteurs de recherche dont nous étudierons l'évolution au contact du TALN[3].

Dans la seconde partie de ce document, nous nous efforcerons de définir quels sont, de nos jours, les problèmes et les besoins des utilisateurs au sein des entreprises pour accéder aux informations pertinentes. Puis, nous mettrons ces besoins en perspective avec les « nouvelles » fonctions des langages classificatoires. Nous nous attacherons plus précisément à la définition et aux applications des ontologies, thesaurus[4], et taxinomies. Enfin, nous dresserons un inventaire des systèmes de gestion de contenu proposant ces applications.

Dans la dernière partie, nous examinerons le projet de création de portail pour le réseau hygiène et sécurité du travail d'Air France. Après une synthèse de l'analyse de l'existant menée pour le service, nous étudierons les taxinomies mises place dans le cadre de l'implantation du moteur de recherche Verity K2 à Air France.

[3] Traitement Automatique du Langage Naturel

[4] On trouve parfois le terme thesauri, qui est le pluriel de thesaurus. Mais, dans les ouvrages en langue française, il est plus courant de voir le terme Thésaurus employé en lieu et place de thesauri..

Partie 1 :

Définitions, outils et contextes de la recherche d'information

1.1 Langages documentaires et recherche d'information

« La vocation de tout système informatique documentaire est de raccourcir et de simplifier le chemin entre l'utilisateur et l'information : toute technologie qui répond à cet objectif doit être intégrée dans le processus de traitement de l'information. » [85 – MANIEZ]

1.1.1 Qu'entend-t-on par information et recherche d'information ?

Karine Lespinasse-Sabourault, dans son cours sur les origines de la linguistique, nous propose cette définition de la recherche documentaire :

« Identification de documents ou d'autres sources d'information ayant des chances d'être pertinentes et utiles par rapport à un besoin d'information permettant à l'utilisateur d'évaluer et si possible d'utiliser ces éléments. » [66-LESPINASSE-SABOURAULT]

Information et document, besoin d'information et recherche d'information, on perçoit dans cette définition, à quel point toutes ces notions sont imbriquées. Toutefois, nous allons tenter de démêler cet écheveau dans la suite de cette étude.

1.1.1.1 Notion d'information et notion de document [9-MANIEZ]

C'est dans la première partie du XX^{ème} siècle, avec le dépouillement des revues que naissent les notions d'information et de document. Les documentalistes, face aux exigences de plus en plus précises des utilisateurs, prennent conscience qu'un ouvrage n'est pas un tout monolithique mais plutôt une compilation de petits blocs informationnels. Chaque bloc correspondant à une approche différente d'un même sujet (analyse sous l'angle économique, sociologique, historique etc.…). Ainsi, l'unité de base documentaire n'est plus le livre ou la revue, mais le chapitre ou

l'article.

- **Définition de l'information**

« L'information est de l'intelligence, du savoir. Une information n'est pas matérielle, elle est abstraite par nature. Pour s'exprimer dans le monde matériel, elle a besoin d'un support. Le document est ce support. L'information est le contenu du document ». (Bernat, L. , Documentaliste, sciences de l'information, vol.32, N°6, 1995, pp 270-282, in [9-MANIEZ p.65] **)**

Toutefois, cette définition correspond à une vision objective de l'information et si l'on s'en réfère à la définition donnée par Didier Frochot, c'est *« uniquement le regard que porte l'être humain sur un objet qui le rend porteur d'information ».* [8-FROCHOT]

Ainsi, tout objet peut devenir un ***document*** porteur d'informations.

- **Définition du document**

Le petit Robert nous donne la définition suivante du document : « Du latin : *documentum : « ce qui sert à instruire »*. Sens actuel issu de l'emploi juridique :

- Ecrits servants de preuve ou de renseignement : annales, archives, documentation, dossier, matériaux, papier, pièce, documents de première main (synonyme : source, information)

- Ce qui sert de preuve, de témoignage : pièce à conviction, enregistrement sonore ou vidéo.

- Pièce qui permet d'identifier une marchandise en cours de transport : factures, polices d'assurance, document administratif unique (D.A.U).

Si l'on se réfère à la définition donnée par Brigitte Guyot, dans son cours d'introduction aux sciences de l'information, un document c'est :

1- Un support (papier, électronique, un objet….)

2- Une inscription selon les règles technico-éditoriales :

- Techniques : logiciel (ex : .pdf, .doc …), formatage informatique (ASCII), écran

- Editoriales :

  - ✓ Forme : image, texte, graphique
  - ✓ Un texte écrit selon une écriture (alphabétique, langue, notation musicale, mathématique (équation), format d'image) et doté éventuellement d'un paratexte

3- Une inscription porteuse de sens (le « contenu »)

4- Muni de marquages documentaires pour le retrouver (accès)

5- - identifiants, liens, classement selon le sens, selon les intentions présumées

Doté de règles socio organisationnelles : statut, cadre d'usage, règles de circulation

Le tout fortement contextualisé : (conditions de productions et de communications).

On le constate aisément, un document peut aussi bien être une œuvre littéraire, qu'un objet ancien, ou un livre de comptabilité, du moment qu'il est porteur d'informations pertinentes pour la personne qui désire le consulter.

C'est pourquoi, il est également indispensable de préciser et de décrire la nature de ce support, car il peut lui aussi être une source d'information pour le chercheur. De même que la facture d'un tableau permet parfois de distinguer l'œuvre de l'élève, de celle du maître, la qualité du papier ou des illustrations d'un manuscrit moyenâgeux, donne des indices sur le statut de son destinataire. Or, ces éléments peuvent être au

cœur d'une recherche documentaire.

Il est donc indispensable, pour référencer un support d'information en tant que « document » dans un système de « stockage » (bibliothèque ou base de données), d'identifier et de décrire : *« ses caractéristiques physiques (sa forme) et ses caractéristiques intellectuelles (son contenu) »*.

Cette opération ne peut se faire au hasard car elle doit répondre aux « processus » intellectuels de la recherche d'information, qu'elle soit réalisée par un homme ou par une machine. En effet, dans un cas comme dans l'autre, il est nécessaire de déterminer les « critères » qui permettent de différencier un document d'un autre afin de pouvoir le retrouver à coup sur. Ils doivent aussi permettre de répondre aux questions que l'on se pose naturellement au moment d'une recherche : Qui ? Quand ? Quoi ? Comment ? Où ?

1.1.1.2 Identifier un document et son contenu

«*La part essentielle de la valeur ajoutée au document par les spécialistes de l'information consiste dans les données représentatives de son sujet »* (Hjorland, Birger, Information seeking and subject representation. An activity-theorical approach to information science, London, 1997, Greenwood Press) in [9-MANIEZ, p.122]

Toutefois, avant même de s'intéresser au contenu d'un document, un grand nombre d'éléments de type « morphologique » permettent de le caractériser. La compilation de ces informations constituent la notice catalographique.

1.1.1.2.1 La description bibliographique

Le titre, l'auteur, la maison d'Edition, le lieu et la date de publication, sont autant de critères de recherche pour un utilisateur. Toutefois, on ne peut décrire un vase de la même manière qu'un texte. Si, certains éléments sont communs : un auteur, une date

de création, un lieu de production, un « titre », d'autres ne sont pas compatibles. Pour une photo, il conviendra par exemple de préciser la nature du support (diapos, argentique) ou le format de la pellicule employée (24x36, Moyen format, etc.).

Par conséquent, pour rendre ce système de catalogage efficace et universel, il a fallu dans un premier temps établir les éléments communs à tout les « supports d'information » susceptibles d'être référencés dans une base de connaissances. Par conséquent la notice minimale doit permettre de renseigner sur le titre du document, son ou ses auteurs, son lieu et sa date de création. Par la suite, en fonction de la nature des documents et de l'usage qui en est fait, la notice peut être complétée par des informations qui permettront de mener une recherche pertinente et plus rapide.

Quelle soit consultable sur papier (catalogue) ou dans une banque de données informatisée, les champs contenant ces informations « tangibles » sont autant de clés d'accès vers le document concerné. Le processus est d'ailleurs le même :

- compiler dans un index le contenu du champ « auteur » de toutes les notices d'un corpus de document,
- classer ces « « auteurs » par ordre alphabétique,
- indiquer pour chaque auteur, la cote des ouvrages

On peut procéder ainsi pour chaque champ, de manière à créer un système d'accès multiples aux informations.

Toutefois, la manipulation et la consultation de cette multitude d'index rendaient la recherche souvent lente et pénible. En automatisant le processus de comparaison et de renvoi entre les index, les premières bases de données informatisées ont apporté une aide précieuse aux utilisateurs pour identifier et localiser les documents dans une collection. Toutefois, ces critères « objectifs » ne constituent qu'une description extérieure du document et ils ne fournissent que très peu d'informations sur son contenu.

En effet, le titre d'un ouvrage n'est pas toujours révélateur de son sujet : comment imaginer qu'un article du magazine « maison coté sud », intitulé : « Au seuil de la

fraîcheur » concerne un Riad au Maroc, plutôt qu'une maison troglodyte ou un recueil de recettes sur les sorbets !!!

On voit ici, à quel point il est nécessaire de fournir la « carte d'identité » complète d'un document. En recoupant les informations concernant la forme et celles concernant le fond, on optimise le processus de recherche.

Toutefois, le sujet d'un livre, ou d'un article est rarement monolithique, et c'est là que réside toute la difficulté lorsqu'on essaye d'en cerner les contours.

1.1.1.2.2 La notion de sujet

Pour Jacques Maniez, c'est une *« Parcelle de l'univers méconnaissable à laquelle s'attache l'intérêt d'un individu »* [9-MANIEZ, p.171]

La complexité de cette notion, tient en grande partie à son ambivalence. En effet, elle se compose de deux aspects indissociables : d'une part, la thématique intrinsèque au document analysé et d'autre part, l'intérêt porté à ce domaine par la personne qui a initié la recherche.

En effet, connaître la motivation du demandeur permet de cibler les types de documents qui apporteront les réponses les plus pertinentes : ouvrages grand public ou scientifique, littérature grise, etc.

Par exemple, lorsqu'un étudiant demande à une documentaliste : « des documents sur le génocide rwandais » ; il est nécessaire pour répondre pertinemment à sa question de savoir quels sont les axes de sa recherche : impact économique, politique, aspects sociologiques, etc…..Par conséquent, pour trouver les « bonnes informations », les documentalistes ne peuvent se contenter de la description « objective » du document : titre, auteur, date de publication, etc…La fiche d'identité de l'ouvrage devra donc comporter des termes explicites, permettant de préciser les notions « subjectives » gravitant autour du sujet principal : le génocide rwandais.

Toutefois, la variété et la « finesse » des thèmes abordés dans un document, peuvent rendre extrêmement complexe l'identification de son sujet « principal ». Décrire le contenu d'un texte ou d'une image, c'est aussi tenir compte de la formulation des requêtes faites par les chercheurs d'information. Il est donc indispensable de concevoir des outils permettant de « normaliser » le vocabulaire utilisé, de manière à le faire coïncider avec les expressions du langage naturel employées par les utilisateurs.

La recherche, qu'elle soit effectuée manuellement ou automatiquement à l'aide d'un logiciel, nécessite une adéquation parfaite entre les termes utilisés pour décrire le sujet et ceux employés pour formuler la requête. Mais, comme nous l'avons vu précédemment la notion de sujet ne peut que difficilement être réduite à une série de « mots-clés », elle doit être construite. Et cette construction doit permettre d'explorer les notions qui entourent le sujet principal du document.

L'opération qui consiste à révéler de façon explicite « les sujets » abordés dans un document, et à les exprimer par l'intermédiaire d'un « vocabulaire » contrôlé s'appelle l'indexation. Cette opération intellectuelle répond à certain nombre de règles que nous allons maintenant évoquer.

1.1.1.2.3 Un processus indispensable : l'indexation

Selon le Robert, indexer c'est « *attribuer à un document une marque distinctive renseignant sur son contenu et permettant de le retrouver ».*

La norme d'AFNOR NF-Z47-102, est même un peu plus précise en ce qui concerne la nature de cette indexation : «L'indexation **humaine** est l'opération qui consiste à décrire et à caractériser un document à l'aide de **représentation des concepts** contenus dans ce document. » [22-AFNOR]

Mais la « description » du contenu d'un document n'est pas une fin en soi, le dictionnaire encyclopédique de l'information et de la documentation (Le Coadic, Nathan, 2001) nous indique que la finalité de l'indexation est de « *faciliter l'accès au*

contenu d'un document ou d'un ensemble de documents à partir d'un sujet ou d'une combinaison de sujets ou de tout autre type d'entrée utile à la recherche ».

Et Suzanne Waller, dans son ouvrage intitulé « l'analyse documentaire », nous rappelle que *« lorsqu'on indexe, on cherche dans le texte des réponses à des questions susceptibles d'être posées*. » [10- WALLER]

L'indexation et la recherche d'information sont donc deux processus intimement liés, qui répondent à une démarche intellectuelle simultanée. Ceci est particulièrement visible lorsqu'on analyse les étapes qui « rythment » l'indexation :

1) **L'analyse documentaire,** qui consiste à reconnaître les concepts essentiels du document :

   - De quoi parle le document ?
   - Comment peut-il être interrogé ?

   L'AFNOR, désigne cette étape comme une « opération qui consiste à présenter sous une forme concise et précise des données caractérisant l'information contenue dans un document, ou un ensemble de documents. Elle permet d'observer, d'identifier et de comprendre un texte avec comme objectif de le rendre utilisable. L'analyse doit être courte et explicite et doit rendre compte du message contenu dans le texte, que celui-ci soit un ouvrage, un article (papier ou électronique), un rapport ou un brevet. »

2) **La sélection des concepts les plus pertinents**

- Les « concepts » retenus doivent être suffisamment développés dans le document analysé pour pouvoir faire l'objet d'une recherche ultérieure.
- Ils doivent correspondre à un **intérêt des utilisateurs** et tenir compte du futur **contexte de consultation** de ces documents : Portail intranet, base de données, index d'un livre, catalogue d'une bibliothèque….En effet, pour un même sujet, les outils, les documents proposés et même les utilisateurs peuvent être différents selon le média utilisé pour accéder aux informations.

En effet, si on prend l'exemple d'un portail Internet consacré à la documentation et accessible au grand public, celui-ci devra impérativement proposer un accès à son corpus de document qui soit « compréhensible » pour les utilisateurs néophytes. Par conséquent, les concepts retenus seront plus généralistes (quittes à guider ensuite l'utilisateur averti vers des notions plus complexes à l'aide d'un formulaire de recherche expert). A contrario, la base de données d'un centre de documentation réservé aux professionnels des sciences de l'information, peut proposer immédiatement des concepts plus complexes et plus détaillés, car ils correspondront à des notions couramment employées par les utilisateurs.

3) **La formalisation de ces concepts en langage naturel**

Elle doit prendre en compte les problèmes :

- D'ambiguïté liés aux « particularités » des langues naturelles
  - Homonymie (une prononciation, plusieurs sens : un seau et un sot)
  - Synonymie (un sens, plusieurs signes : gérer / administrer
  - Polysémie (un signe, plusieurs sens : la marmotte qui est un animal pour les biologiste et accessoire pour les opticiens)
- De « variabilité » des formes lexicales :
  - Substantifs
  - Adjectifs
  - Verbes
- De structure des unités lexicales :
  - Unitermes : agir
  - Mots composés : entrer en action

4) La traduction de ces concepts en langage documentaire

« *Si l'indexation des documents est une opération complexe et incertaine, les langages documentaires ont été crées pour la simplifier, la guider et la normaliser »* [9- Maniez p.169]

La première fonction des langages documentaires est donc d'« éliminer » les problèmes intrinsèquement liés à la nature « variable » du langage naturel. C'est pourquoi, les langages documentaires sont souvent désignés sous le titre de «vocabulaire contrôlé ».

Par conséquent, l'opération consiste à choisir le ou les descripteurs (ou les indices, mots-clés, et autres vedettes - matières selon le langage documentaire employé), qui permettent d'obtenir la meilleure « transcription » du concept retenu.

En représentant le contenu d'un document originel (primaire) sous une forme condensée (liste de descripteurs ou résumé), le documentaliste produit un document secondaire, plus facile à consulter. Par ailleurs, avant l'apparition des outils de Geide et des systèmes de recherche en texte intégral, ce processus représentait le seul moyen pour obtenir un bref aperçu du contenu d'un document. Aujourd'hui, c'est l'abondance des fichiers numériques qui rend la consultation difficile, ce qui permet au document secondaire de (re)trouver son utilité première en apportant une aide à la sélection.

L'ensemble de ces opérations aboutit à l'indexation du document. Il ne reste plus maintenant qu'à évaluer la qualité de cette indexation. Si cette procédure n'a rien de systématique, il est néanmoins intéressant pour le documentaliste de procéder régulièrement à des tests qui lui permettront de vérifier la qualité des opérations effectuées. En effet, une mauvaise indexation (incompréhension du texte, omission de certaines thématiques) génère systématiquement du « bruit » ou du « silence », lors de la recherche. Toutefois, une question mal posée peut avoir les mêmes conséquences et nécessiter une reformulation.

C'est pourquoi, toute réflexion sur les processus d'indexation, doit prendre en compte les principes qui régissent la recherche d'information. Ces deux démarches sont intimement liées, et l'on ne peut tenter d'améliorer l'une sans que cela ait des conséquences sur l'autre.

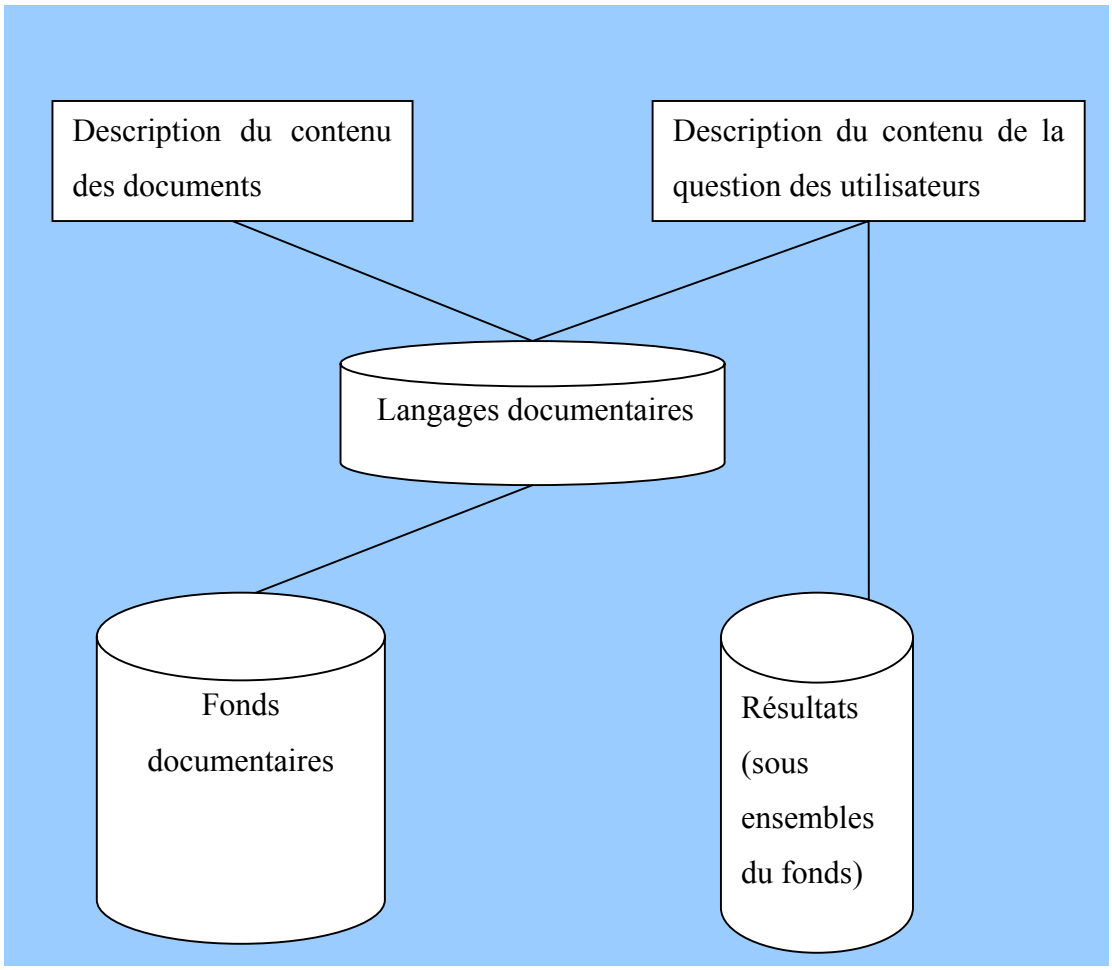


Figure 1: les langages documentaires : des « langages pivots » au coeur de la recherche d'information . Boydens Isabelle- 2 avril 2004 – Du Web sémantique au web pragmatique – SmalS-MvM

1.1.1.3 Besoin et recherche d'information

« Le succès de la recherche exige que le système et l'usager expriment les clés dans les mêmes formes, exigence que ne respecte pas la liberté créative des langues naturelles… » [9-MANIEZ, p.170]

1.1.1.3.1 Quelques définitions

Tout système de recherche d'objets (papier ou informatisé) est conçu en vue *« de sélectionner rapidement des objets spécifiques à la demande des usagers* [9-MANIEZ, p.170]

Dans notre cas, l'objet recherché est de l'information (textuelle, chiffrées, sonore, etc.…). Or, on entend fréquemment parler de : recherche *d*'information, de recherche de ***l'***information et de recherche ***documentaire.*** Le flou terminologique qui entoure ce processus nécessite une petite « mise au point » avant de poursuivre notre analyse.

Le « Vocabulaire de la documentation ».- ADBS Ed : Paris, 2004, nous propose les définitions suivantes :

Recherche d'information :

Ensemble 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

Recherche documentaire :

Ensemble des 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.

Recherche de l'information :

Ensemble des méthodes, procédures et techniques ayant pour objet d'extraire d'un document ou d'un ensemble de documents les informations pertinentes.

On le constate, une gradation existe entre ces trois formulations. De la recherche d'information « au sens large » dans un corpus de document non « limités », on passe au repérage et à la sélection des documents susceptibles de constituer un réservoir d'informations pertinentes, pour finir par l'extraction de ces informations dans un « lot » de documents présélectionnés.

Chacune de ces procédures représentent-elles une « stratégie » adaptée à un certain type de question ou constituent-elles les étapes successives d'une recherche complexe ?

Or, on recherche de l'information pour quelqu'un ou pour soi-même. Par conséquent, l'utilisateur, ses besoins, la façon dont il les exprime, jouent un rôle prépondérant dans la stratégie de recherche à adopter pour trouver l'information pertinente.

« La pertinence peut être définie comme l'adéquation de l'information aux besoins de celui qui la reçoit » [16-ACCART]

En effet, même si le « sujet » de la recherche est identique, les besoins en information d'un décideur, d'un médecin ou d'un technicien ne seront pas les mêmes, car ils n'en font pas le même usage. Par conséquent, les informations fournies et la manière dont elles sont communiquées seront très différentes d'une personne à l'autre. Ainsi, au sein du réseau hygiène et sécurité du travail d'Air France, on observe que les demandes d'information (concernant les accidents du travail) en provenance du personnel de terrain, concernent le plus souvent des points techniques ou juridiques, nécessitant des réponses rapides et concises, pour résoudre un problème urgent. En revanche, les instances dirigeantes du réseau et de la compagnie demandent plus fréquemment des évaluations chiffrées du taux d'accidents correspondant à tel ou tel risques, une estimation des impacts de la campagne de prévention mise en place ou des comparatifs avec les mesures prises dans les autres entreprises, etc. Dans ce type de demande, les informations

rassemblées, sont destinées à fournir une aide décisionnelle aux cadres de l'entreprise. Par conséquent, elles doivent être actualisées en permanence (ce qui suppose un système de veille), et « synthétisées » de manière à ce que les utilisateurs puissent en prendre connaissance de manière rapide. Ce « traitement » de l'information fait partie intégrante de la mission du documentaliste et il n'est pas sans influence sur la sélection des documents.

1.1.1.3.2 Méthodologie de la recherche [16-ACCART]

On observe une méthodologie commune à tous ces types de recherche :

1) Identifier le besoin réel de l'utilisateur, si nécessaire reformuler sa demande
2) Définir le champ de la recherche et déterminer la stratégie à mettre en œuvre pour y répondre
3) Identifier les sources documentaires existantes
4) Extraire les informations essentielles des sources identifiées
5) Evaluer les résultats obtenus et réajuster éventuellement la démarche.

La première étape est tout particulièrement importante, car elle conditionne tout le reste du processus. En effet, le plus souvent les questions posées par les utilisateurs restent assez vagues et n'abordent que la thématique « générale » de leur recherche. La mission de la documentaliste consiste alors, à aider l'utilisateur à formuler le contexte et la problématique de sa recherche, pour lui permettre d'échapper à une exploration « linéaire » du plan de classement.

En effet, si certaines questions trouvent une réponse immédiate dans la consultation d'un document ou d'un article de dictionnaire, d'autres nécessitent de rassembler de l'information en explorant une grande quantité de documents. Or, ces sources d'informations sont rarement rassemblées dans une même catégorie thématique.

En décomposant la problématique principale de la requête, en plusieurs « sous - problèmes », le « chercheur d'information »[5] peut les réduire à des concepts, plus faciles à combiner pour interroger des systèmes de recherche d'objet (base de donnés, site intranet, etc.).

On voit à quel point, la formulation de la question a une influence sur la méthode employée pour, mener à bien la recherche.

1.1.1.3.3 Recherche factuelle, recherche documentaire, recherche contextuelle : quelles différences ?

L'ouvrage de J.P Accart : « le métier de documentaliste », nous propose **une typologie des questions** posées par les utilisateurs. Nous allons tenter de déterminer, à quels modes de recherche elles correspondent. [5-ACCART]

La question ponctuelle :

« Elle appelle une réponse unique et immédiate : une date, une adresse, le nom d'un auteur etc. »

Elle induit une **recherche factuelle**, qui correspond plus à une demande de renseignement qu'à une recherche d'information. Elle doit répondre à des questions précises du type : « Donnez moi la date de naissance de Louis XIV ? » ou « Quel est le nom du réalisateur de Citizen kane ? ». La réponse à fournir devient une « parcelle » de l'information concernant un sujet plus vaste (pour les exemples cités : la vie de louis XIV ou l'œuvre d'Orson Wells).

Les premiers systèmes d'information susceptibles de répondre à ces questions, sont les dictionnaires, et autres répertoires proposant des « clés » d'entrée par ordre alphabétique. Le sujet principal de la recherche étant, en général, ramené à un seul mot, il est facile de lui trouver une correspondance dans l'index de l'encyclopédie consulté.

[5] Les outils s'étant démocratisés, il devient parfois difficile d'utiliser systématiquement le terme documentaliste, car les « amateurs » sont de plus en plus souvent confrontés aux mêmes situations que les professionnels.

En proposant un index pour chaque critère de description (champ), et en offrant la possibilité de réaliser des recherches croisées, les bases de données informatisées ont nettement faciliter les recherche factuelles. En effet, ce principe permet de multiplier les clés d'accès et de les combiner afin de restreindre la recherche au document le plus pertinent (celui qui les contiendra toutes). Parmi, les représentations les plus emblématiques de ce type de systèmes de recherche, nous trouvons le centre de renseignements d'SVP et le site des pages jaunes (www.pagesjaunes.fr).

Toutefois, Jacques Maniez, dans son ouvrage sur les langages documentaires, soulève une autre forme de questions ponctuelles. Celles dont la formulation induit une clé d'accès *« trop fine ou trop complexe pour qu'une encyclopédie la sélectionne comme point d'entrée »* et qu'il qualifie de **recherche contextuelle**. En effet, il est souvent impossible de répondre à une requête du type « qui est l'auteur de … ? », car aucune encyclopédie papier ne propose dans son index une entrée, par le titre d'un ouvrage ou le début d'une citation. [9-MANIEZ]

Dans cette situation, le recours aux systèmes de recherche en « full text » (texte intégral), devient beaucoup plus « efficace » pour les chercheurs d'information. Ces moteurs, en utilisant des index alimentés par tous les mots d'un champ ou d'un texte, se montrent *« capables de cibler un contexte marqué par l'occurrence (ou mieux, la coocccurence) des termes d'une question, [ils] offrent une solution de secours en cas de défaillance des banques de données factuelles. »* [9-MANIEZ, p.111]

Ainsi, si on cherche l'auteur des « Malheurs de Sophie », en tapant sur Google « Malheurs » et « Sophie », on obtient 47 000 réponses, parmi lesquelles on trouve des documents concernant la série télé du même nom ou des articles sur les récentes mésaventures de l'actrice Sophie Marceau à Cannes !!! Pourtant, et malgré tout ce « bruit », les deux premières références de la page apportent l'information demandée : « ouvrage écrit par la comtesse de Ségur ».

On le constate aisément, cette méthode de recherche tient beaucoup au hasard, même si elle « répond » (plus ou moins pertinemment) à un certain besoin d'information. Toutefois, son meilleur contexte d'utilisation reste Internet grâce à la facilité avec laquelle on peut formuler les requêtes, ce qui le rend beaucoup plus accessible au

grand public.

- **La question chronologique** :

« Qui fait intervenir la notion de déroulement et a donc un caractère rétrospectif : par exemple, l'histoire d'un peuple de telle période à telle autre. »

- **La question de synthèse**

« Qui demande au documentaliste de rassembler des références permettant à l'utilisateur de faire la synthèse des connaissances acquises au jour de la demande. »

- **La question du type « état de l'art »**

« Où le documentaliste ne sélectionnera que les documents d'actualité sur des recherches ou des projets en cours. »

Contrairement à la question ponctuelle pour laquelle on attend une réponse tenant en un mot, ces trois derniers types de question impliquent des réponses argumentées et documentées. Elles nécessitent la collecte et la consultation de plusieurs sources d'information, afin de traiter la demande de l'utilisateur sous tous les angles correspondant au champ d'investigation défini avec lui. [5-ACCART]

Ainsi, même dans le cas d'une question chronologique, du type : « l'histoire du Rwanda depuis l'indépendance », le documentaliste ne pourra pas se contenter de fournir un seul ouvrage (aussi bon soit-il) à son commanditaire. Le cas échéant, il prendrait le risque de ne proposer qu'une réponse « orientée », voir tronquée si le livre est un peu ancien. Par conséquent, pour ce type de question, l'adjonction de cartes géographiques commentées et/ou d'articles de presse traitant des évènements récents est une nécessité.

C'est pourquoi, dans ces trois cas, le documentaliste doit procéder à **une recherche documentaire**, c'est-à-dire à une identification des références bibliographiques pertinentes. Cela lui permet d'avoir un aperçu global de la littérature produite sur ce sujet, et de découvrir parfois, des pistes de réflexion insoupçonnées. Ainsi, le

documentaliste peut ensuite « piocher » les documents les plus à même de satisfaire à la demande de son commanditaire.

La difficulté de la recherche documentaire, c'est-à-dire par « sujet », réside dans le processus de collecte des références des documents. Contrairement, à la recherche factuelle, les clés d'interrogation de la recherche documentaire, ne se résument pas à un ou deux termes « concrets[6] » (en général un nom de personne ou de lieu). En effet, la description du « sujet » d'un document correspond à une combinaison « personnalisée » de clés. Pour pouvoir le retrouver, le documentaliste doit reconstruire cette association de concepts (ou s'en approcher au plus près). Cette démarche permet de rétablir le contexte sémantique des clés employées, ce qui est absolument indispensable pour éviter de retomber dans les « travers » liés à la nature du langage naturel : polysémie, homonymie, etc.…

Comme on progresse dans l'arborescence d'une classification, du plus général au plus précis (ou inversement) ; on construit une requête en partant du terme qui illustre au mieux le domaine étudié, puis on lui associe d'autres concepts qui permettent de restreindre ou d'élargir le périmètre de la recherche[7]. Mais, ce processus pour être efficace nécessite de pouvoir « visualiser[8] » l'outil documentaire (thesaurus, classifications, etc…) qui a permis de réaliser l'indexation du corpus exploré. Sans cela, la combinaison des mots-clés devient une juxtaposition de termes « contrôlés », mais « décontextualisés », générant à la fois « bruit » et « silence ».

Pendant plusieurs années, face aux progrès des systèmes de recherche statistiques et linguistiques, la tendance était à délaisser les langages documentaires, devenus moins « pratiques » pour indexer les documents. Tout au plus, l'éditeur se contentait-il de fournir quelques listes d'autorité pour faciliter l'interrogation de certaines bases de données. Aujourd'hui, face au déluge informationnel provoqué par Internet (19,2 milliards de pages indexées sur yahoo), les éditeurs s'intéressent à nouveau à ces « organisateurs » de la connaissance. Mais, avant de nous pencher sur les nouveaux

[6] Jacques Maniez emploie le terme de discret, c'est-à-dire bien défini.
[7] Dans le cas des thésaurus intégré à une base de données informatisée, cette fonction prend le nom d'autopostage.
[8] Même si le thésaurus est invisible car intégré dans le système de recherche, il faut au minimum que le logiciel propose les concepts en relation avec le terme inscrit dans le champ de recherche. Dans le cas contraire il prive l'utilisateur de pistes d'investigation, ce qui va à l'encontre de la fonction d'un thésaurus.

usages des langages documentaires, il est nécessaire de procéder à quelques petits rappels historiques et fonctionnels.

1.1.2 Qu'est ce qu'un langage documentaire ?

Selon l'AFNOR, ce sont : « des langages artificiels faits de représentations de notions et de relations entre ces notions, destinés, dans un système documentaire, à formaliser les données contenues dans les documents et les demandes des utilisateurs. » [22-AFNOR]

1.1.2.1 Historique

- **A l'origine**

Outils incontournables de la recherche d'information, depuis des décennies les langages documentaires se sont tout d'abord imposés pour faciliter l'archivage des documents.

Les premiers systèmes de classement, créés pour organiser le contenu des bibliothèques, s'inspiraient du fonctionnement naturel de l'esprit humain, qui instinctivement, trie, classe et hiérarchise les informations qu'il perçoit selon des critères allant, en principe, du général au particulier.

Ainsi, les premières bibliothèques rangeaient leurs ouvrages par taille (in quarto, in octavo) ou par ordre alphabétique des auteurs, à l'intérieur de grandes thématiques correspondant aux disciplines universitaires telles que les sciences naturelles, ou la philosophie. (B. Beghtol, Clare – les domaines de la connaissance : Multidisciplinarité et système de classification bibliographique, 1998 – 12 pages).

Puis, face à la multiplication des livres et des documents conservés dans les bibliothèques, il est apparu nécessaire de décrire le contenu des ouvrages afin de mieux répondre aux demandes des utilisateurs. Pour mener à bien ce processus d'indexation, il était nécessaire d'établir au préalable une « cartographie » des connaissances. A chaque branche du savoir correspondait plusieurs classes et sous-

classes, subdivisées jusqu'à l'obtention d'une couverture exhaustive des sujets traités par les ouvrages de la bibliothèque. Cette opération en « normalisant » le vocabulaire utilisé et les relations hiérarchiques entre les thématiques permettaient à ces classifications de cumuler trois fonctions : «la *représentation du contenu intellectuel des documents (indexation), recherche des références documentaires pertinentes (fichiers systématiques matières) et accès aux documents eux-mêmes (cotation) »* [20-FEYLER]

Jusqu'à la fin des années 80, les classifications hiérarchiques, nées au 19ème siècle, étaient encore le moyen d'accès aux documents, le plus répandu. Toutefois, leur structure extrêmement lourde à modifier, présentait l'inconvénient de cloisonner les ouvrages dans une seule thématique. Ces contraintes se révélaient particulièrement handicapantes pour les professionnels de l'information, qui devaient compter essentiellement sur leur mémoire pour localiser les ouvrages susceptibles de traiter d'un sujet transversal.

- **L'âge d'or des langages combinatoires**

Peu à peu, les classifications décimales[9] ont été délaissées au profit de langages permettant de décrire plus finement le sujet d'un document, ce qui devenait indispensable pour faciliter la recherche documentaire. Mais, dès cet instant, deux questions incontournables s'imposaient :

- Comment décrire le contenu d'un document, sans que l'affect de la personne en charge de cette tâche ne vienne « perturber » l'opération ?
- Comment adapter le vocabulaire d'indexation à la « culture » des utilisateurs ?

C'est en conservant ces deux questions à l'esprit, que les documentalistes commencèrent à élaborer des outils documentaires permettant de décrire le sujet des documents, « *non plus globalement, mais par une combinaison de concepts* ». [15-CONTAT]

[9] Aujourd'hui, elles sont principalement utilisées pour leur fonction de cotation.

La forme la plus simple de ces langages analytiques est la liste d'autorité. Elle est constituée d'une série de termes « normalisés » (en général masculin singulier), définis selon leur contexte d'utilisation. A chaque terme de cette liste, sont associés des synonymes de manière à proposer d'autres clés d'entrée pour les utilisateurs.

Plus complexes et très répandus dans les années 90, les thésaurus sont composés de descripteurs structurés en catégories et reliés entre eux au moyen de relations sémantiques hiérarchiques ou associatives. Ces « descripteurs » choisis en commun par les bibliothécaires, constituent en quelque sorte une « cartographie sémantique » de la collection. En proposant un accès par les « non descripteurs », c'est-à-dire les synonymes des termes retenus, le thésaurus permet un accès plus « naturel » aux documents. Par ailleurs, l'existence de relations, du type « voir aussi », entre les descripteurs représente un bond considérable dans le domaine de la recherche documentaire. Cette opération, ancêtre du lien hypertexte, offre l'opportunité de découvrir simultanément les ouvrages correspondant au cœur de votre recherche et ceux susceptibles de compléter ou d'apporter un éclairage nouveau à votre sujet principal.

Toutefois, la manipulation (physique et intellectuelle) des thésaurus restait souvent trop complexe pour les non initiés. Or, avec l'extraordinaire développement d'Internet depuis les années 1990, les utilisateurs ont acquis une autonomie dans la recherche d'information, qui s'accorde mal avec le recours systématique à un documentaliste (ce qui s'impose avec les versions papier des thésaurus !!!)

- **La révolution informatique**

Jusqu'au milieu du XX^{ème} siècle, les fonctions des langages documentaires connurent peu d'évolutions et restèrent principalement dédiées à l'indexation et à la localisation (cotation) des documents. En revanche, en quelques années, l'apparition de l'informatique entraîna de profonds bouleversements dans le domaine de la recherche documentaire.

On peut situer les véritables débuts de l'informatique documentaire entre 1965 et 1972, parallèlement à l'explosion des services de recherche et développement des groupes Industriels. A l'époque, l'informatique documentaire est encore

principalement considérée comme une aide à l'archivage. Les premiers systèmes mis en place prennent la forme de banques de données, dans lesquelles sont archivées les notices catalographiques des documents conservés sous format papier dans les centres de documentation. Si ces outils permettent uniquement une recherche par champs (y compris sur les mots-clés), ils nécessitent néanmoins la création d'un langage spécifique pour procéder à la requête[10].

A cette époque, le grand public (encore sous équipé en matériel informatique) n'a pas accès à ces langages d'interrogation réservés aux professionnels. Les spécialistes de l'information, ayant parfois, eux-mêmes, recours à un informaticien à qui ils transmettent la requête « reformulée » de l'utilisateur. Ce processus d'interrogation, décentralisé et en différé (batch), disparaîtra rapidement face à la démocratisation des outils bureautique. [20-FEYLER]

A partir des années 80, on voit apparaître des systèmes de recherche de plus en plus pointus utilisant les techniques de l'indexation automatique et la recherche en texte intégral. Pendant quelques années ces technologies vont continuer à cohabiter avec les langages d'indexation, d'interrogation et de classement, censés être devenus inutiles….. Ce partenariat se traduisait le plus souvent par une introduction des « descripteurs » dans les index constitués automatiquement par le logiciel à partir du recensement de tous les mots du texte.

- **Le choc Internet**

Avec l'explosion d'Internet au début des années 90, les interfaces de recherches du web, basées sur une coordination booléenne implicite[11] entre les termes de l'index, ont rencontré un rapide succès auprès du grand public. D'un abord plus facile que les langages d'interrogation, elles ont engendré un « oubli » des fonctionnalités des langages documentaires. Toutefois, face à l'augmentation croissante de l'information disponible sur le WEB, les concepteurs ont très vite pris conscience que cette technologie générait beaucoup de « bruit » mais aussi du « silence ».

[10] A l'image du langage, encore utilisé de nos jours pour interroger la base Questel.
[11] Dans les moteurs de recherche développé pour le web, les termes juxtaposés sont automatiquement associé par l'opérateur « et ». Le grand public est rarement informé, qu'il a la possibilité de remplacer ce « et » par d'autres opérateurs, tel que le « or ».

A l'époque, le traitement automatique de l'indexation ne fonctionnant que sur des principes statistiques, il ne permettait pas de prendre en compte les problèmes d'homonymie, de synonymie et de contextualisation, liés à l'emploi du langage naturel par les utilisateurs. Ces derniers, peu formés à l'emploi des opérateurs booléens, n'étaient pas en mesure de construire les requêtes complexes, nécessaires à la « réduction » du périmètre de recherche[12]. Les spécialistes de l'information, conscients des limites « binaires » des systèmes statistiques, travaillèrent en collaboration avec les informaticiens et les linguistes, à la création de systèmes capables de reproduire les capacités d'analyse et d'interprétation du cerveau humain. Aux moteurs de recherche en texte intégral, les chercheurs ajouteront peu à peu des notions de linguistique, puis de sémantique dans le but de créer des systèmes de gestions de contenus en langage naturel, de plus en plus pertinents. [32-SERRES]

Ces recherches sur l'Intelligence Artificielle sont encore loin d'être achevées, toutefois elles ont permis de faire évoluer la perception et les applications des langages classificatoires. Plus proches du fonctionnement de l'esprit humain, dont le premier reflex est toujours de trier, classer et archiver les informations qu'il perçoit (même si cette opération se fait de manière inconsciente), les classifications et les thésaurus sont apparus comme de « nouveaux outils » pour organiser les informations en amont et en aval de la recherche. On vit alors apparaître chez les éditeurs de logiciel et dans la littérature spécialisée, une vague de nouveaux outils : ontologies, taxinomies, topics, etc….Mais que signifie réellement ces termes, quelles sont les fonctionnalités effectives de ces outils ? Sont-ils réellement nouveaux ?

Avant de pousser plus loin notre analyse à ce propos, il m'est apparu nécessaire de dresser une typologie des langages documentaires « classiques », pour mieux cerner le rôle et l'origine de ces « nouveaux outils ».

[12] Pour une recherche sur « les nouvelles autoroutes de l'information », si on tape « autoroute » et « information », nous obtiendrons énormément de documents sur la diffusion de l'information pour les automobilistes empruntant l'autoroute. En revanche, si on tape : « autoroutes » et « information » et « internet » sauf « automobile, on augmente la quantité de résultats pertinents.

1.1.2.2 Inventaire et typologie des langages documentaires

Les langages documentaires constituent un ensemble d'outils sémantiques riches et subtils. Les fonctions de ces langages sont si proches et si variables selon leur contexte d'utilisation (papier ou numérique) que les chercheurs ont multiplié les tentatives de constitution d'une typologie « idéale ». Ainsi, certains opposent les langages synthétiques (utilisant une syntaxe) aux langages énumératifs (sans syntaxe), d'autres les langages d'interrogation aux langages d'indexation et de classement. Certain, comme Jacques Maniez, propose une structuration fondée sur la notion « de sujet et les modalités de sa représentation ». Dans cette dernière, les langages sont répartis en trois familles : [9-MANIEZ]

- **La Famille thématique** qui correspond à une approche territoriale du sujet. Le sujet des documents est « identifié » à l'aide d'une grille thématique pré-établie et hiérarchisée. Elle comprend, par conséquent les langages de classifications (CDU), tel que les taxonomies.

- **La famille « syntagme** » : qui permet de traduire très fidèlement, « *en une formule univoque, à la fois les termes et la structure syntaxique du syntagme représentant le sujet* ». Ce groupe correspond aux langages à facettes crées par Ranganathan. Les ontologies qui permettent d'établir des relations du type « est la conséquence de », « est mangé par », peuvent être rattachées à ce groupe (voir Partie 2)

- **La famille « concept »** : qui consiste à définir le sujet d'un document, en délimitant « sa surface référentielle par un nombre restreint de concepts bien choisis ». Le membre le plus représentatif de cette famille est le thésaurus.

On remarquera que les deux dernières familles sont représentées par des langages combinatoires, fonctionnant respectivement selon un mode linguistique[13] et booléen.

On retrouve, par conséquent, le « clivage » traditionnel qui divise les langages documentaires en deux groupes : Les langages combinatoires et les langages classificatoires.

Cette typologie traditionnelle est celle reprise par l'AFNOR, qui propose la terminologie suivante :

- **LANGAGES HIERARCHISES** :

  - **Classification** : « langage documentaire fondé sur la représentation structurée d'un ou plusieurs domaines de la connaissance en classes et dans lequel les notions et leurs relations sont représentées par les indices d'une notation ». [22-AFNOR, p.39]

  - **Nomenclature** : « classification méthodique de l'ensemble des termes d'un domaine spécialisé » [22-AFNOR, p.84]

  - **Plan de classement (ou plan de classification)** :

    « *Structure hiérarchique et logique permettant (la classification), le classement et le repérage de pièces d'archives ou d'ensembles documentaires. (...) Il peut être : a) général s'il permet de regrouper tous les documents d'un service d'archives; ou b) spécifique s'il permet de répartir les documents d'un seul fonds ou collection.* » (Archives nationales du Québec, 1996, p. 154). C'est le reflet matériel d'une classification intellectuelle.

  - **Taxonomies :** A la fin du XVIIIème siècle, ce terme désignait principalement les classification hiérarchique des espèces vivantes. On retrouve d'ailleurs la même racine que dans Taxidermie. Comme cette

[13] C'est à dire syntaxique, ex « les châteaux en France, « en France » est le complément de lieu de château, (relation symbolisée par le code R7) , par conséquent la formule de recherche de ce sujet sera « château R7 France », plus précis que Château « et » France, qui pourrait donner des résultats concernant l'influence de la France sur les bâtisseurs de château dans le monde.

dernière, la taxonomie conserve une « image », une mémoire du « vivant » à un instant T. Nous verrons plus en détail, dans la seconde partie de ce document, les applications actuelles des taxonomies dans le milieu de l'informatique documentaire.

- **LANGAGES COMBINATOIRES** :

  - **Lexique** : liste des mots d'une ou plusieurs langues dans un domaine donné [22-AFNOR, p.73]

  - **Liste d'autorité** : liste des vedettes ou termes qui doivent être obligatoirement et nécessairement utilisés dans le catalogage ou l'indexation [22-AFNOR, p.74]

  - **Thésaurus** : langage documentaire fondé sur une structuration hiérarchisée d'un ou plusieurs domaines de la connaissance et dans lequel les notions sont représentées par des termes d'une ou plusieurs langues naturelles et les relations entre notions par des signes conventionnels [22-AFNOR, p.112]

  - **Liste de vedettes-matières :** Ensemble d'un ou plusieurs descripteurs exprimant et précisant le sujet d'un document. Chaque vedette-matière correspond à un seul sujet, simple ou complexe. Un même document peut avoir plusieurs sujets donnant lieu à la rédaction de plusieurs vedettes matières. (AFNOR, 1996b, NF Z 44-070, p. 441)

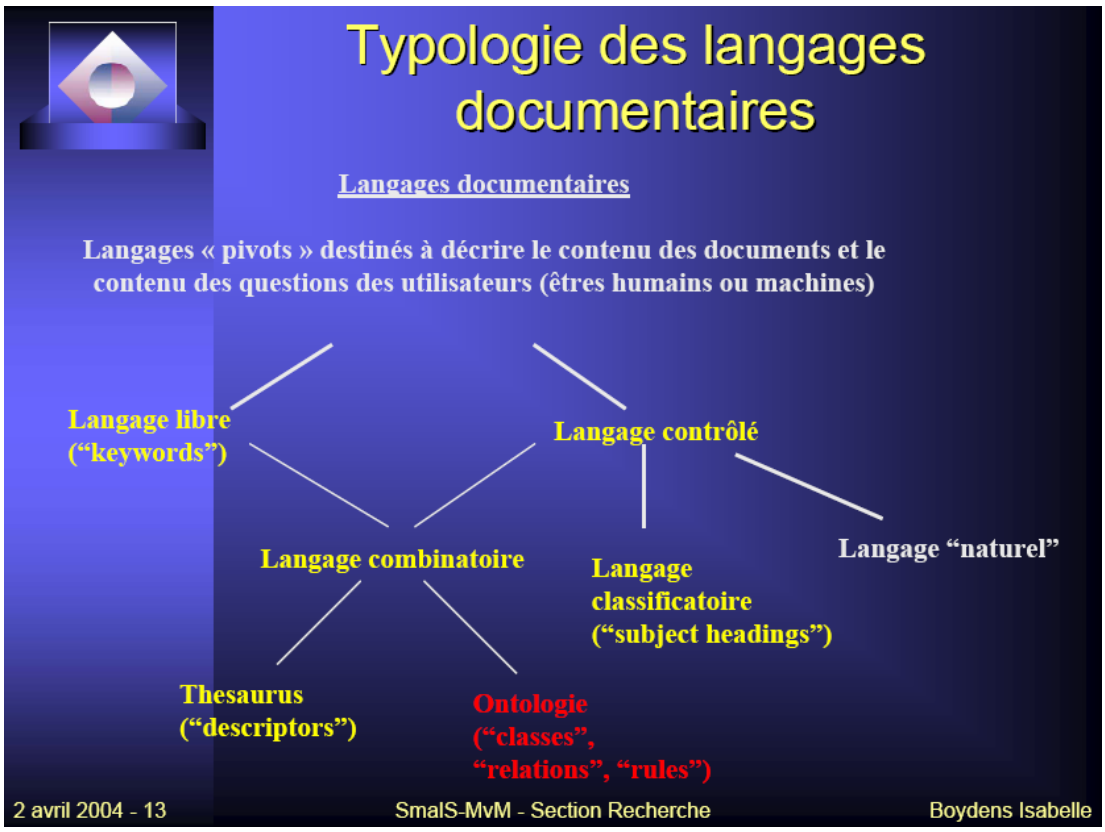


Figure 2 : Typologie des langages classificatoires. Boydens Isabelle

Nous verrons dans la seconde partie de ce document, que cette typologie peut être quelque peu remise en cause lors du transfert de ces outils sur le web. Si nous prenons l'exemple du thésaurus, ce dernier est traditionnellement un langage d'indexation et d'interrogation. Importé dans un site web, ce dernier devient un outil de navigation qui peut s'apparenter à un plan de classement. En circulant dans les champs sémantiques du thésaurus web comme dans une arborescence, l'utilisateur affine peu à peu sa recherche et accède finalement aux documents pertinents, et pas seulement à leurs notices.

1.1.2.3 Inventaire et historique des classifications

« En structurant le savoir et le savoir-faire en disciplines et en domaines pratiques parcellisés et hiérarchisés, nos civilisations créent des cadres classificatoires de fait, qui contribuent à baliser le territoire des sujets matières » [9-MANIEZ, p.131]

Il ne s'agit pas ici d'évoquer les « nouvelles » applications des langages classificatoires, mais plutôt de rappeler les origines de ces outils et leurs fonctionnalités « classiques ». Leur évolution face à l'émergence des systèmes de gestion informatisés et les progrès de la recherche sur les outils en langage naturel, sera traitée dans la seconde partie de ce mémoire.

1.1.2.3.1 Quelques rappels terminologiques

A l'origine des langages classificatoires, il y a les classifications :

Michèle Rive nous propose plusieurs définitions qui nous éclairent sur leur nature et leurs fonctions [14-RIVE]

« *Le classement, c'est l'action de ranger effectivement d'après un certain ordre »*

« La classification, c'est l'ensemble des règles qui président au classement effectif ou qui déterminent idéalement un ordre dans les objets ».

« Elle s'établit des termes les « extensifs » ou génériques aux plus « intensifs » ou spécifiques. »

« Hiérarchiser, c'est donner un ordre (un ordonnancement) à des objets que l'on a déjà distingués entre eux. »

« La catégorisation se fait par rapprochement, comparaison avec des caractéristiques déjà connues et ordonnées »

Mais, comme nous l'avons vu dans la partie précédente une classification « **documentaire »** (dans son acception traditionnelle), est plus qu'une simple

organisation hiérarchisée des concepts, ordonnant les connaissances du général au particulier. Sa spécificité consiste à codifier ces concepts sous la forme de chiffre (codification numérique) ou de chiffres et de lettre (codification alphanumérique) permettant de « localiser » l'emplacement physique où le document est conservé. Ce sont les **INDICES**.

1.1.2.3.2 Chronologie des classifications « historiques »

1876 La classification de Dewey

Elle est divisée en 10 classes de 0 (généralités) à 9 (histoire, géographie) subdivisées décimalement en divisions et en sections. C'est la première à introduire une notation décimale pour représenter les différentes classes et sous-classes. Plus le chiffre est long, plus on descend dans la hiérarchie.

1897 La classification de la bibliothèque du congrès de Washington

C'est une classification pragmatique (elle s'inspire directement du contenu de la collection) comportant 23 classes, dont 2 sur l'histoire de l'Amérique.

1906 La classification décimale universelle de Paul Otlet et Henri Lafontaine

C'est un index général de classement couvrant toutes les connaissances pour tous les hommes et tous les temps. Elle comporte 10 classes et 33 000 subdivisions, son index comporte 38 000 entrées. Sa grande nouveauté consiste dans l'utilisation combinée de chaînes numériques et de symboles. Ils sont au nombre de deux :

- l'extension (une barre oblique « / ») : elle signifie qu'on inclue toutes les notions comprises entre les deux éléments séparées par cette barre.
- la relation (deux points « : ») : ce signe introduit, au contraire de l'autre, un rapport entre les deux notions qu'il sépare.

Pour la première fois, les concepts choisis pour représenter les connaissances ne sont plus perçus comme des éléments isolés, se suffisant à eux-mêmes, mais comme des sujets liés les uns aux autres. L'apparition de ces relations de type « sémantique »[43]

entre les catégories, représente un grand progrès pour les utilisateurs, qui peuvent ainsi découvrir des éléments tout à fait pertinents pour leurs recherches mais non directement liés à celle-ci. [23]. Cette « fonctionnalité » est particulièrement utile dans les bibliothèques spécialisées où l'emploi de classes très détaillées, et décrites séparément est fréquente.

On ne peut que constater à quel point la frontière est proche entre ce type de classification et les thésaurus, qui utilisent eux aussi un système de relation entre les « descripteurs ». La différence tenant essentiellement au rôle dévolue à chacun de ses outils : indexation et aide à l'interrogation pour le thésaurus, organisation physique et intellectuelle d'un fond pour les classifications.

1933 La classification de Colon (Ranganathan)

Classification dite « analytico-synthétique », qui associe à chaque domaine de la connaissance (histoire, philosophie, médecine, etc.) 5 concepts fondamentaux qui représentent les « facettes » de ce domaine, c'est-à-dire les points d'entrée contenus dans le « sujet » de la requête [9-MANIEZ], [11-MANUE]:

- Personnalité : c'est l'objet dont il est question.
- Matière : qui représente la substance (la propriété) de cet objet
- Energie : qui représente l'opération (l'action) en rapport avec cet objet
- Espace : le lieu où prend place cet objet
- Temps : la période, l'époque où existe cet objet

« La classification à facette est une classification non énumérative, disponible sous formes de tables de concepts regroupés en classes homogènes. Dans chacune des classes, les concepts sont divisés en groupes connus sous le nom de « facettes ». Au sein de chaque facette, les termes représentant les concepts peuvent être classés de façon hiérarchique. » (Michèle Hudon, Sabine Mas, 2001, p.6).

Les classifications à facettes offrent une infinité de combinaisons entre les « sujets de bases » et les « sujets secondaires » d'un document, ce qui permet sa description

« sur mesure » et une pluralité des clés d'accès. A contrario, dans une classification énumérative (Dewey), la structure hiérarchique étant fermée, elle offre une seule voie d'accès, très précise mais limitée à une thématique principale.

Ainsi, dans l'exemple proposé par Jacques Maniez [9-MANIEZ , p.221] :

Pour un sujet sur « la prévention des maladies Virales du riz », la classification Dewey propose :

633 = Céréales
633.18 = Riz
633.189 = maladies du riz
633.189.8 = maladies du riz d'origine virale

On voit que la notion de prévention ne peut être prise en compte dans la description du sujet (de la requête ou de l'indexation) car elle est absente de la classification.

A l'inverse, la CC[14] propose :

Facette principale (EJ) = Agriculture.
Facette personnalité (381) = RIZ.
Facette Matière (421) = Maladie virale.
Facette Energie (5) = Prévention.

De ce point de vue, l'indexation proposée par la CC est celle qui traduit de la manière la plus complète, le sujet de la requête exprimée en langage naturel. Toutefois, son utilisation dans les phases d'interrogation est tellement complexe, qu'elle a été peu allégée voir délaissée en faveur d'autres classifications en particulier celle de DEWEY.

[14] Classification de Colon

1.2 Systèmes de gestion de contenus et outils de la RDI

« Un système d'information est donc avant tout un système d'acteurs intervenant autour d'objectifs, d'un projet commun, que ce soit de façon ponctuelle ou permanente ; il a des moyens, et a mis en place des procédures. Il peut ne pas y avoir un système technique, mais il y a toujours un système organisationnel. » [79-GUYOT,p.17]

« [Les systèmes de recherches d'objets] *comportent un module de stockage et un module d'extraction. Le module de stockage organise les objets en vue de leur sélection : mise en ordre de la collection ; affectation des attributs adéquats à chaque objet (clés de recherche) ; mise en mémoire et en ordre de ces attributs. Le module d'extraction prévoit un mode d'interrogation du système à partir des mêmes clés, ce qui implique un dialogue entre le système et l'usager. » [9-MANIEZ, p. 17]*

Dès l'apparition des premières bases de données informatisées, l'indexation et le catalogage des documents sont les processus, qui ont « cristallisé » toutes les angoisses des spécialistes de l'information et des créateurs de logiciel. Nous l'avons vu, l'indexation résulte d'une opération intellectuelle individuelle, or les langages documentaires mis à la disposition des documentalistes ne peuvent que limiter la « variabilité » entre deux indexeurs (culture, niveaux de vocabulaire différents, maîtrise de l'orthographe). En aucun cas, ils ne peuvent éliminer les erreurs ou les différences d'appréciation inhérentes à la nature humaine (omissions, fautes de frappe, mauvaise compréhension d'un mot-clé, etc.).

Or, ces erreurs de transcription peuvent avoir des conséquences, non négligeables, sur un processus de recherche informatisée. En effet, là où un humain est capable de rétablir naturellement l'orthographe correcte d'un auteur qu'il connaît, par exemple : *Enstein* à la place d'Einstein, le logiciel, lui, les considèrera comme deux auteurs différents. Par conséquent, lorsqu'une requête sera effectuée avec l'orthographe correcte de l'auteur, les ouvrages mal référencés ne pourront apparaître. Pour améliorer ce processus, les informaticiens en collaboration avec les documentalistes ont mis en place des « aides à la saisie » : en permettant un

« remplissage » direct de certains champs grâce à des listes d'autorité (« faites maison » ou achetées comme RAMEAU pour le champ « description »). Toutefois, cette technique ne peut pas être utilisée pour tous les champs (ex : le titre) [15] et elle n'est réellement efficace que si le « chercheur d'information » a lui aussi, accès à la liste d'autorité (les erreurs de saisie ont autant de conséquences dans la phase d'interrogation que dans la phase d'indexation.)

Avec la dématérialisation des fonds documentaires, liée à l'explosion des documents « numérités[16] », ce problème est devenu encore plus sensible.

Pour ce type de support, un mauvais référencement équivaut à un « aller simple » pour les oubliettes. En effet, contrairement aux livres d'une bibliothèque, les fichiers numériques (référencés dans une base de données) sont abrités sur un serveur distant interdit d'accès aux utilisateurs « lambdas ». Par conséquent, ils ne peuvent être consultés « par hasard » et doivent faire l'objet d'une requête pour être mis à disposition d'un lecteur. C'est pourquoi, depuis plusieurs années, les éditeurs de logiciel portent une attention particulière aux technologies visant à améliorer le catalogage et l'indexation de ce type de document.

L'un des principaux axes de recherche a consisté à développer des structures (invisibles pour l'utilisateur) qui permettent à l'auteur de signaler au logiciel de recherche les éléments constituant la « carte d'identité » de son document. Conçues pour décrire et naviguer entre les documents publiés sur le Web, les langages HTML [17] (Hypertext Markup Language) et XML[18] (eXtensible Markup Language), permettent d'intégrer directement dans le texte des balises qui structurent l'information qu'ils contiennent.

« Ces informations sur (préfixe *Meta*) l'information » constituent les « méta données ». Nous verrons plus précisément, dans la seconde partie de ce mémoire, quels sont les apports de ces langages pour les logiciels de gestion de l'information.

---

[15] Le titre d'un ouvrage est unique et ne peut être réutilisé pour remplir d'autres notices.
[16] Néologisme, crée par Fabrice Molinaro, qui en donne la définition suivante : « Contrairement, à la numérisation qui effectue une copie d'un document déjà existant, la numérité définit l'état d'un document qui n'existe que sous format numérique. »
[17] Voir définition p.69
[18] Voir définition p 69

Si ces deux « nouveautés » ont permis d'optimiser l'indexation et le référencement des documents, elles n'ont pas pour autant supprimées l'intervention humaine et les « risques » qui en découlent.

Alors, comment lutter contre ces « incidents » typiquement humains ?

En automatisant complètement l'indexation ? En rendant les logiciels de gestion de contenu plus « intelligents » ? Ces deux voies ont été explorées et elles ont donné naissance à des outils de plus en plus perfectionnés que nous allons étudier dans la seconde partie de ce mémoire.

1.2.1 Qu'est ce qu'un système de gestion de contenu ?

« Information électronique, Internet, etc. : il n'y a plus de différentiation. Ces éléments forment un tout, « l'information numérique », qui correspond mieux à la réalité » [41-REMIZE, p.18]

Dans le domaine informatique, la gestion de contenu couvre tout le cycle de vie des documents disponibles dans le système d'information de l'entreprise. Ils prennent en compte l'ensemble des informations circulant dans l'entreprise, c'est à dire :

- Les documents statiques stockés dans des systèmes de fichier (contenus non structurés).
- Les documents dynamiques (textes, images, son, etc…) mis à disposition sur le web, et en provenance des bases de données de l'entreprise (contenus structurés).

Ces deux types d'information correspondent aux deux domaines de la gestion de contenu :

- La GED traditionnelle
- La gestion de contenu Web ou CMS pour *Content Management System*.

Dans les deux cas, le cœur du système repose sur un référentiel qui permet d'organiser les informations de l'entreprise en les associant à des métadonnées qui permettent de les identifier et de les localiser dans le système. Ils incluent aussi des processus de workflow qui établissent les modalités de création, de validation et de circulation des documents au sein des organisations. La particularité des CMS est de proposer autour de ces fonctionnalités centrales, un ensemble de prestations qui permettent d'harmoniser le traitement et la diffusion des informations et documents. Ainsi, en amont, ils peuvent proposer des logiciels pour créer des documents, et en aval, des infrastructures (Portail) pour mutualiser les informations et des outils permettant de les « retrouver » (moteur de recherche).

1.2.2 Inventaire et typologie des outils CMS.

« Aujourd'hui, une partie de l'information numérique est créée en anticipant à la fois sur les façons dont on va la rechercher dans le futur et sur la façon de la stocker. » [41-REMIZE, p.18]

Il y a quelques années encore, la typologie « classique » distinguait les outils de recherche selon **le mode de recherche** qu'ils proposaient :

- **Par navigation** :
  - Dans une arborescence (annuaires, classifications)
  - Par liens hypertextes (sites, base de signets)
- **Par requête** :
  - Moteurs de recherche :
    « Programme qui indexe le contenu de différentes ressources Internet, et plus particulièrement de sites web, et qui permet à l'internaute de rechercher de l'information selon différents paramètres, en se servant de mots clés, et d'avoir accès à l'information ainsi trouvée. » [16-ACCART, p.404]

- Assistants ou fédérateurs de requêtes
  « *Ces logiciels traduisent les requêtes des utilisateurs dans les différents langages d'interrogation propriétaires ou dans les protocoles d'interrogation des bases de données en ligne* [22-AFNOR]

Mais, aujourd'hui on assiste à une hybridation de plus en plus fréquente des outils. Ainsi, la plupart des outils de navigation par arborescence dissimulent des liens hypertextes derrière leurs descripteurs et permettent ainsi un accès direct aux documents. Parallèlement, les moteurs de recherche intègrent dans leur fonctionnement interne des arborescences[19] et des index (constitués à partir des mots du texte et des descripteurs), qui sont employés pour affiner les requêtes.

Par conséquent, cette typologie n'est plus vraiment à l'ordre du jour, en revanche d'autres critères d'indexation semblent aujourd'hui plus pertinents.

Le premier est apparu en même temps que l'informatique documentaire et se fonde sur **le mode d'indexation des ressources** :

- **L'indexation humaine,**
  qui concerne principalement les annuaires thématiques.

- **L'indexation automatique,**
  principe de fonctionnement des moteurs de recherche statistiques.

Malgré la fusion de plus en plus fréquente des annuaires et des moteurs de recherche, cette typologie reste d'actualité car ces outils ne se substituent pas l'un à l'autre. Ce sont avant tous des outils complémentaires, dont le rôle, l'usage et la technologie sont différents. En effet, un annuaire classe (et par conséquent indexe) les sites dans des thématiques prédéfinies, ce qui permet d'opérer un repérage des sources d'information potentiellement intéressantes. Mais il ne permet pas d'avoir une vision détaillée de leur contenu.

[19] Par exemple : Yahoo pour Internet est à la fois un annuaire et un portail.. Quant au moteur de Verity pour les portails intranet, il propose une recherche guidée grâce à des taxinomies.

En revanche, un moteur de recherche, en indexant toutes les pages d'un site, permet d'accéder directement aux documents correspondant à la recherche. Toutefois, ces « bonnes » réponses sont souvent noyées dans une multitude de résultats parfois sans rapport avec l'objet de la requête. C'est pourquoi, l'annuaire thématique en guidant l'utilisateur en amont de la recherche, permet d'améliorer la pertinence des moteurs en éliminant les problèmes de polysémie, liés au fonctionnement des systèmes de recherche en texte intégral.

Alexandre Serres (maître de conférence en Sciences de l'information à l'URFIST de Rennes), propose **une autre typologie** présentant l'avantage d'intégrer de nouveaux outils recherche, ignorés par les précédentes répartitions.

Elle repose sur « **le mode de fonctionnement interne** » des outils :

- **Les « outils de première main »** comportant leurs propres bases de ressources d'information et un module d'interrogation dédié. Cette « famille » comprend :

  - **Les moteurs de recherches**

    Ce sujet fait l'objet d'un développement plus approfondi dans la partie consacrée aux portails et aux moteurs de recherches, nous vous renvoyons vers la p.58 de cet exposé.

  - **Les annuaires ou répertoires :**

    Ces outils sont en quelque sorte des bases de données rassemblant et organisant thématiquement des références de sites. Chaque site est décrit dans une notice comme un livre peut l'être dans un catalogue (informatisé ou pas). Les utilisateurs peuvent explorer l'annuaire à l'aide d'une arborescence dans laquelle les sites peuvent se répéter autant que nécessaire selon le nombre de thématiques qu'ils abordent. Sur le web, les annuaires sont systématiquement couplés à des moteurs de recherche permettant d'effectuer une exploration par mots-clés. Les utilisateurs obtiennent ainsi, une liste de sites susceptibles de contenir les informations qu'ils recherchent. *« un répertoire permet donc d'aider à repérer les sites existant dans un domaine d'activité, et à ce titre d'aider à la cartographie*

du web » [37-FROCHOT, p.2]

- **Les « outils de seconde main »** comportant uniquement un module d'interrogation et qui exploite les bases des moteurs et des annuaires. Ils comprennent :

  - **Les métamoteurs** :

    « Outil qui, pour une même requête, interroge en parallèle (simultanément) plusieurs moteurs de recherche et/ou répertoires et compile les résultats avant de les présenter. Parfois qualifiés d'agents semi-intelligents, de métachercheurs…. » [16-ACCART, p. 404]

    A part Copernic dont l'usage est destiné au « grand public », les métamoteurs sont principalement utilisés en Interne par les grandes entreprises, pour *« filtrer, analyser, cartographier, résumer …des corpus de documents hétérogènes »* [32-SERRES, p.4]

  - **Les portails** :

    « Terme générique pour désigner un site qui sert de point d'entrée sur l'Internet ou Intranet pour un nombre important d'utilisateurs. Un site portail offre une multitude de contenus, outils et services différents depuis la page d'accueil. Certains sont des répertoires professionnels organisés autour d'une thématique particulière, dotés d'un moteur de recherche interne et d'autres ne sont que des points d'entrée sur l'Internet offerts par les fournisseurs d'accès. » [16-ACCART, p. 404]

  - **Les outils de veille ou outils intelligents.**

    « Ces outils intelligents sont destinés à effectuer des comparaisons de chiffres et d'informations, des analyses de documents, etc…Ce sont des outils capables d'autonomie, de collaboration (avec d'autres outils), et d'adaptation à son environnement. » [32-SERRES, p.4]

  - **Les outils annexes**.

    On range sous cette appellation vague un ensemble d'outils diversifiés

pouvant servir à la recherche d'information et à la veille de manière annexe : « aspirateurs de sites web, organisateurs de signets …. »

Ce mémoire s'attachant à définir les nouvelles fonctionnalités des langages classificatoires dans les systèmes de recherches informatisés, et leurs apports pour les utilisateurs des portails Intranet ; nous allons étudier plus en détail le mode de fonctionnement et l'évolution des Portails et des moteurs de recherche. Ainsi, on note depuis quelques années, l'émergence d'un clivage qui tend à opposer outils généralistes et outils spécialisés. Si cette opposition ne remet pas en cause les typologies précédemment citées, elle introduit néanmoins de nouveaux paramètres qui ont eu un impact important sur les fonctionnalités de ces deux outils.

1.2.3 Portails Intranet et moteurs de recherche

Ces deux outils sont aujourd'hui les moyens les plus répandus pour stocker, gérer et transmettre l'information sur Internet. Rapidement, les entreprises se sont emparées de ces technologies pour organiser l'accès aux documents publiés sur leurs réseaux intranet, Mais, pour être efficaces ces outils doivent être adaptés à l'information qu'ils accueillent et au public auquel elle est destinée.

1.2.3.1 Qu'est ce qu'un portail ?

1.2.3.1.1 Définition générale

« *Un portail Internet* [ou intranet] *est un site qui offre la possibilité de disposer, à partir d'un point d'accès unique, de données issues de sources multiples. Il permet d'accéder rapidement à une information qualifiée, organisée et structurée, personnalisée en fonction des centres d'intérêt des usagers* ». [34 –CHARVET]

Avec le succès d'Internet et la multiplication exponentielle des sites Web, il s'est rapidement avéré nécessaire d'organiser l'accès aux informations et aux services proposés sur la toile. Ce rôle a été dévolu aux portails qui ont du s'adapter à la demande du public en fournissant des services de plus en plus spécialisés. Par

conséquent, nous avons vu apparaître plusieurs familles de portails aux fonctionnalités bien définies.

1.2.3.1.2 Typologie des portails

- **Les portails généralistes**

  Ils sont destinés au « grand public », c'est à dire à une multitude de « profils » utilisateurs, qui représentent autant de centres d'intérêt. Le rôle du portail est de faciliter et d'organiser l'accès à ces contenus informationnels, de manière à ce que tous les publics puissent localiser les informations qu'ils recherchent dans les meilleurs délais. Par conséquent, ces portails sont le plus souvent munis d'un annuaire de sites (organisé selon une classification thématique « généraliste ») et un moteur de recherche en texte intégral du type Google. En parallèle, ils proposent un certain nombre de service (météo, flash actualités, programme culturel de la semaine, etc…) destinés à fidéliser les utilisateurs potentiels, qui pourront ensuite, faire de ce portail leur page d'accès par défaut au réseau. Free et Yahoo font partis de cette famille de portail, fournisseurs d'accès à Internet, ils proposent de surcroît des accès sécurisés à des boites de messagerie personnelle, des « lieux » d'échanges communautaires (yahoo group) et des espaces de publications web personnel (pagesperso, blog). La tendance actuelle est la personnalisation de ces portails, l'utilisateur a la possibilité de « customiser » la page d'accès en sélectionnant les rubriques et services qui correspondent à ses besoins.

- **Les portails thématiques ou spécialisés**

  Fondamentalement, leur principe de fonctionnement est le même que celui des portails généralistes, pour ce qui est de l'accès aux informations. En revanche, ils ne proposent pas ou rarement des services du type messagerie ou météo, mais plutôt des prestations en corrélation avec la spécialisation et les attentes de leur public : newsletter, forums de discussions en rapport avec la spécificité du site, offres promotionnelles, etc…En effet, suivant que le site est destiné aux professionnels (portail du BTP), au grand

public (portail sur le cinéma, le droit, le sport, etc…) ou à un public d'amateurs avertis (portail du surf, portail des métiers d'art), les services proposés et leur mode de diffusion seront différents. [84 – ACCART]

- **Les portails des portails**

  La multiplication des portails spécialisés a naturellement appelé la création de « super » portail destiné à organiser et mutualiser les adresses des autres portails présent sur le web, au sein d'un annuaire ou par l'intermédiaire d'un moteur de recherche. Cette tendance se retrouve sur les intranets des grands groupes internationaux.

- **Le portail d'entreprise**

  Il peut être une simple « vitrine » proposant la carte d'identité de la société ou un véritable lieu d'échanges, où l'internaute peut procéder à des achats en ligne, consulter le catalogue de produits ou les archives ouvertes de l'organisation, dialoguer avec certains services (après-vente), ou accéder aux offres de recrutement. Certains de ces portails offrent même un accès sécurisé au réseau intranet du groupe, pour le personnel habilité. C'est le cas d'Air France[20], ce qui est particulièrement utile aux PNC et PNT[21] qui ne disposent pas de bureau au sein de l'entreprise de pouvoir se connecter de chez eux ou lors des escales, au réseau via le portail : Intralignes (http://intralignes.airfrance.fr).

  De même qu'il existe une typologie des portails Internet, il existe aussi une typologie des réseaux intranet. Elle tient principalement à la nature et aux fonctions que chaque organisations leurs assignent, ce que nous allons étudier maintenant plus en détail.

[20] Pour plus de facilité, j'utiliserai parfois le sigle AF pour désigner la compagnie
[21] Personnel navigant technique et commercial, c'est à dire les pilotes et les hôtesses ou stewart.

1.2.3.1.3 Les portails Intranet des entreprises : des informations variées destinées à un public hétérogène

Contrairement aux portails Internet qui donne accès à un stock de documents ouvert, constamment renouvelé, destinés à un public non défini, les portails Intranet sont destinés à gérer des lots de documents validés, correspondant aux activités de l'entreprise et consultables par un public professionnel.

Toutefois, la cohabitation de plusieurs corps de métiers au sein d'une même entreprise, implique des besoins informationnels et fonctionnels différents, ce qui a entraîné la création de trois grandes « familles » de réseau intranet.

Du point de vue technique, un Intranet est un :

« *Réseau informatique interne à une entreprise ou à une institution, ...[qui] ne peut être consulté qu'en local. »* . [16-ACCART, p. 404]
Ce réseau utilise « *tout ou partie des technologies et des infrastructures de l'Internet pour transporter et traiter les flux d'informations internes d'un groupe d'utilisateurs identifiés.* » [46-ALERI, p. 45]

Les réseaux Intranet répondent à 3 grandes exigences des entreprises :

- ✓ La transversalité de l'information
- ✓ La mutualisation des applications informatiques
- ✓ L'homogénéisation des accès à l'information et aux outils applicatifs

Selon l'activité, la taille et les infrastructures de l'entreprise déjà en place, certaines de ces exigences prendront le pas sur d'autres, et détermineront la nature du réseau mis en place :

- **Un intranet documentaire** :
  ils sont destinés à organiser la production, la publication et l'administration des collections documentaires. La qualité de ce type de réseau réside dans l'efficacité et la qualité du référencement des

documents et des modes de recherches proposés aux utilisateurs (taxinomies, recherche en texte intégral, interface personnalisée, glossaire métier, etc…). Ils sont tout particulièrement indiqués pour les entreprises où cohabitent plusieurs directions indépendantes et dispersées géographiquement, ce qui induit une redondance des bases documentaires et une mauvaise connaissance des documents produits par chaque entité. La réussite de ce type d'intranet repose sur la mise en place de processus à la fois souples et rigoureux, permettant de « valider » la production et la publication des documents (workflow).

Dans la phase de production, il est capital de définir les critères qui permettront d'identifier chaque document dès leur création et de les gérer tout au long de leur vie au sein de l'entreprise : auteur, mot-clés, durée de vie, n° d'archivage, N° de la version, nom du destinataire etc….Ces données peuvent soient être saisies dans une base de données (logiciel de Geide) et/ou être intégrées directement dans le document (XML : eXtensible Markup Language).

- **Un intranet applicatif**

  « *La migration d'applications centrales ou client/serveur traditionnelles vers un système intranet se justifie donc souvent car cela permet d'une part de profiter des avantages liés au déploiement et d'autres part de bénéficier d'un niveau de standardisation ouvrant sur l'interconnexion de services applicatifs* » [46-ALERI, p. 121]

  L'intérêt de ce type d'intranet est de proposer un accès centralisé à des outils ou à des sources d'informations produites par un service mais pouvant être utile à tout ou partie des salariés de l'entreprise : planning du personnel, formulaire de congés, banques de données internes, offres de mobilité interne, veille stratégique et économique, règlement interne, outils de workflow, groupware, etc…

- **Un intranet d'intégration**

  Il est destiné à « *fédérer les applications hétérogènes existantes, par l'intégration des progiciels et des développements spécifiques dans un environnement graphiques homogène* » [46-ALERI, p. 134]

> Cette démarche permet d'offrir aux usagers un accès homogène aux applications proposées par l'entreprise, quel que soit le point d'accès utilisé par le salarié. Son profil, géré à partir de l'annuaire d'entreprise et identifiable grâce à système d'accès sécurisé, lui permet de bénéficier d'un « bureau », toujours identique, quel que soit le lieu où l'ordinateur qu'il utilise. Il offre en outre l'avantage de faciliter la gestion et la mise à jour des logiciels informatique (nouvelles versions de Word, mise à jour des anti-virus, etc…). Ces applications étant hébergées dans un lieu unique, le serveur, leurs « maintenance » n'imposent une intervention individuelle sur chaque poste de l'entreprise.

A un moment ou un autre, tous ces besoins organisationnels ont été successivement ou simultanément ressentis au sein des grandes entreprises. Au point de provoquer un développement anarchique des bases de données et autres outils applicatifs, sur les réseaux Internes des sociétés.

Face à cette situation, les entreprises ont adopté le principe du portail, pour fédérer et structurer l'accès aux sources informationnelles publiées sur ces réseaux. Mais sans de bons outils de recherche, adaptés à la structure fonctionnelle de l'entreprise et aux besoins des utilisateurs, un portail devient une simple page d'accueil enrichie de quelques services, rien de plus. C'est pourquoi, ce type de CMS est systématiquement fourni avec un ou plusieurs systèmes de recherches complémentaires, le plus répandu étant le moteur de recherche en texte intégral.

1.2.3.2 Les moteurs de recherche en texte intégral

Définition : « En anglais *search engine*, un moteur de recherche est un outil qui recense automatiquement des ressources se trouvant sur internet. Par opposition aux répertoires, un moteur offre une base de données référençant des pages en texte intégral » [32-FROCHOT]

Ils « *sont en quelque sorte le maillon le plus avancé de l'informatique documentaire, qu'on préfère nommer […] informatique du contenu* [36-FROCHOT]

Comme tous les CMS, les moteurs de recherche en texte intégral stockent leurs données dans un fichier maître. Mais, contrairement aux systèmes accueillant des données structurées, réparties en plusieurs zones appelées champs, les systèmes en texte intégral sont destinés à gérer des fichiers non structurés où le titre du document et le reste du texte constituent les seuls champs du fichier.

Les systèmes en texte intégral utilisent des fichiers inversés (index) crées autour du fichier maître pour accélérer le processus de recherche[22]. Cet index recense l'ensemble des mots de chaque champ, soit dans le cas du texte intégral, la totalité des termes présents dans le document[23].

Avec la généralisation des documents « numérités », les moteurs de recherche en texte intégral sont devenus indispensables pour traiter les informations contenues dans les corpus électroniques. Afin d'améliorer leurs performances, ces moteurs intègrent des techniques d'analyses statistiques et linguistiques de plus en plus performantes.

Ces technologies interviennent, dans des proportions variables, au cours des deux étapes fondamentales qui marquent le fonctionnement de tous les moteurs de recherche :

- ✓ L'extraction et l'indexation des termes qui « identifient » le document,
- ✓ La comparaison ou *matching* de ces termes indexés avec ceux de la question posée.

Selon les techniques mobilisées lors de ces deux processus, la qualité et la quantité des résultats obtenus peuvent varier de manière très significative. C'est pourquoi, à chaque « famille » de moteurs correspond un environnement d'exploitation adapté.

[22] Ce qui permet de classer les informations en ordre alphabétique ou numérique, système qui facilite la recherche en évitant de procéder à une exploration séquentielle.

[23] Nous verrons ultérieurement que les avancées linguistiques ont permis d'insérer des « dictionnaires de mots vides », permettant d'exclure des index des termes tels que : de, et, le, un etc….

1.2.3.2.1 Moteurs de recherche à dominante statistique

Comme leur nom l'indique, « *ils s'appuient sur des méthodes statistiques qui leur permettent de répondre à une requête en fonction de la fréquence d'apparition des concepts demandés et de leur répétition au sein de chaque document* » [45-CROCHET-DAMAIS]

Ces méthodes consistent à « guider » l'indexation Full Text, en fonction :

- ✓ De la fréquence,
- ✓ De la position
- ✓ Et de la pondération des mots dans le texte.

Quant à la phase de recherche, elle se fonde sur « *des méthodes de calcul statistiques basées sur l'occurrence et la co-occurrence des mots dans le texte, comparés à sa fréquence dans le corpus.* ». Ce qui revient à dire, que les algorithmes statistiques permettent de reconnaître la présence de chaînes de caractères identiques, dans les divers documents d'un même corpus, et de leur accorder un « poids » (une importance) plus ou moins significative, suivant leur fréquence d'apparition[24].

Pour obtenir des résultats fiables et suffisamment fins, les algorithmes doivent traiter une grande masse d'informations. Par conséquent, les moteurs statistiques sont plus fréquemment utilisés dans des portails généralistes (altavista, google) ou dans des bases de presse, comportant une très grande quantité de documents hétérogènes. Pour la recherche d'information sur les Intranet, le moteur statistique le mieux placé sur le marché est encore aujourd'hui Autonomy. Celui-ci compte parmi ses clients, la société Reuters, qui fournit des informations boursières en temps réel aux professionnels et au grand public (chaînes télé spécialisées).

[24] Principe du moteur Autonomy, basé sur la théorie de Shannon et les algorithmes bayesiens, ou plus un terme est rare dans un corpus, plus son poids est important.

1.2.3.2.2 Moteurs de recherche linguistiques et sémantiques : les apports du TALN

Définition du TALN : « *Le Traitement automatique des langues (TAL) est une discipline à la frontière de la linguistique et de l'informatique, qui concerne l'application de programmes et techniques informatiques à tous les aspects du langage humain. Parmi les applications les plus connues, on peut citer : la traduction automatique (historiquement la première application, dès les années 1950) ; la correction orthographique ; la recherche d'information ; le résumé automatique ; la génération automatique de textes, la synthèse vocale, la reconnaissance de la parole »* [74]

Les moindres performances des moteurs statistiques sur des corpus de documents spécialisés (par conséquent moins volumineux), les rend peu attractifs pour des entreprises comportant des « réseaux métiers » très hétérogènes. Dans ce contexte, certains termes professionnels peuvent prendre une « connotation » différente d'un corps de métiers à l'autre, alors que d'autres peuvent être omniprésents, sans pour autant permettre une indexation significative. Ainsi, dans la liste de mots-clés proposée par la photothèque d'Air France, le terme avion ne renvoie à aucune image, puisqu'il peut être associé à toutes !!!! L'activité d'une compagnie aérienne étant centrée sur les avions, on les désignera plutôt par leur modèle (Boeing 737, airbus A320, ect…). De même qu'on ne parle pas du « personnel travaillant dans l'avion ou hors de l'avion », mais de personnel navigant (P.N) en opposition au personnel sol (P.S).

De plus au sein des entreprises, les ambiguïtés sur le langage peuvent être fréquentes. Or, les moteurs statistiques ne peuvent résoudre ce type de problème sans l'introduction de techniques linguistiques (et sémantique). Si on reprend l'exemple d'Air France, le terme sécurité peut être utilisé dans le sens « sûreté » : lutte contre le terrorisme, le vandalisme, mais aussi dans le sens « qualité de l'entretien des avions » et enfin dans le cadre de la sécurité du travail, c'est à dire la réglementation des conditions de travail des salariés. C'est pourquoi, il est nécessaire de prendre en compte le contexte sémantique des termes utilisés dans les entreprises, en intégrant le

« sens du langage » au processus de recherche.

Ceci ne peut se faire qu'en introduisant des couches de traitements linguistiques dans le fonctionnement des moteurs et en délimitant le périmètre thématique du corpus grâce à des dictionnaires ou des thésaurus spécialisés.

En conséquence, on passe d'une indexation en « texte intégral », qui ne fait qu'isoler des chaînes de caractère séparées par des blancs, à un véritable système de recherche et d'indexation en « langage naturel ». Or, nous l'avons vu le langage naturel est, par essence, soumis à de nombreux problèmes d'ambiguïté, de variabilité et de structuration des unités lexicales.

Pour venir à bout de ces difficultés, les textes à indexer sont traités et analysés en plusieurs étapes :

1. **Pré traitement linguistique :**
   Il se déroule en deux temps **:**
   - La tokénisation (ou segmentation) qui consiste à diviser les textes en unités lexicales (token) de plus en plus petites : paragraphes, phrases, mots. C'est une opération de type statistique qui « localise» les chaînes de caractères entourées de « séparateur » (caractère blanc), et les identifie comme étant des mots.
   - Cette opération est couplée à un traitement linguistique qui permet d'identifier les signes de ponctuation séparant les phrases et les paragraphes. Il permet aussi de procéder à une première correction des fautes d'orthographe[25] et des erreurs de typographie.

2. **Traitement morpholexical et morphosyntaxique :**
   Cette première opération que l'on nomme aussi « lemmatisation », est indispensable pour pouvoir « *retrouver tous les documents dans lesquels apparaissent différentes formes du même mot* » [33-IVANCIUC DENIAU, p.66]. Elle consiste à faire correspondre les formes des termes rencontrées dans les textes (féminin, masculin, adjectifs, verbes, adverbe, substantif) à

[25] Elles seront intégrées dans un « dictionnaire », de manière à ce qu'une requête formulée avec ces mêmes erreurs, puissent malgré tout être reconnu par le moteur et aboutir à une réponse pertinente.

leur « LEMME », c'est à dire la forme fixe et minimale (canonique) du mot.

Le second traitement appelé aussi « étiquetage » ou *tagging* consiste à comparer chaque mot du texte (susceptibles d'être ambiguë), avec les termes du dictionnaire intégré (référentiel ou glossaire métier). Ceci, afin de leur attribuer une ou plusieurs étiquettes en fonction du sens qu'ils sont susceptibles d'avoir dans le contexte où ils sont utilisés. Cette opération permet aussi d' « identifier » les mots composés et les expressions toutes faites.

3. **Le traitement sémantique général de nature lexical**

   Il consiste à identifier les réseaux sémantiques[26] qui unissent les concepts en présence dans le corpus indexé. Ces réseaux constituent un graphe (topic map[27]) qui sert de référence pour l'indexation du fonds et peut ensuite servir de « système de guidage » pour l'utilisateur au moment de la requête[28]. En associant les termes présents dans le corpus par « famille », cette opération permet de diminuer les problèmes de silence et de bruit, liés à la synonymie, à l'hyperonymie (meuble/siège), la métonymie (partie de) ou l'association.

4. **Traitements statistiques**

   A ce stade, les moteurs linguistiques réintroduisent des opérations de type statistiques (cf. p.60), qui permettent de pondérer les termes retenus précédemment.

5. **Traitement de regroupement / classification**

   Ces deux opérations consistent à rapprocher les documents similaires en les classant dans des thématiques (catégories) en fonction de leurs degrés de pertinence par rapport à la question posée[29].

---

[26] « La constitution de ces réseaux passent souvent par l'utilisation de dictionnaires spécialisés qui pour chaque terme lemmatisé donne la distance (proximité) aux autres termes. Ainsi le terme Bébé, appartient-il au réseau sémantique : enfant, garde, nourrice, jouet, mère…etc. » (www.memodata.com).

[27] De topic qui veut dire sujet en anglais et map qui signifie carte.

[28] Le moteur de recherche Internet Kartoo, fonctionne selon ce principe.

[29] C'est à dire leur degré d'appariement calculé. (documentaliste sciences de l'info, vol.37, n°5-6, p. 34

On parle alors de classification automatique si l'opération est effectuée en mode non supervisé par le documentaliste, et de catégorisation automatique, lorsque l'opération nécessite un contrôle humain.

Dans les deux cas, le moteur doit s'appuyer sur des outils de classement issus des langages classificatoires traditionnels mais adaptés aux fonctionnements actuels des ordinateurs : ontologies, taxonomies, thesaurus, etc.

Les apports de ces nouveaux outils, vont bien au-delà de leurs fonctions de classement. Dans la seconde partie de ce mémoire, nous allons étudier les enjeux de l'information au sein des organisations. Puis, nous dresserons un inventaire détaillé de ces outils et de leurs fonctionnalités, en observant la façon dont ils contribuent à améliorer l'accès aux documents, mis à la disposition des salariés au travers des portails d'entreprises.

Partie 2 :

Nouvelles fonctionnalités des langages classificatoires : de l'accès à l'information à la navigation informationnelle.

2.1 Une évolution qui tente de répondre aux besoins des utilisateurs

« En milieu ouvert, c'est-à-dire lorsque le médiateur doit servir une population hétérogène (hétérogénéité des métiers au sein d'une institution, diversité des interlocuteurs dans le cas d'un public tout venant), l'efficacité du transfert de l'information va reposer en grande partie sur sa capacité d'adaptation «aux autres » (au « répertoire » des autres) » [86-SUTTER, p. 87]

Avec Internet l'utilisateur est devenu un « inconnu » possédant milles visages, qui accède directement aux outils de recherche. Par conséquent, le documentaliste ne peut plus jouer son rôle de médiateur qui reformule la question, précise le contexte et adapte la sélection des documents en fonction du profil de l'utilisateur. Sa mission a évolué. Il est là pour concevoir des outils qui vont remplir cette fonction de médiation en structurant les accès à l'information de manière à personnaliser la recherche individuelle. Si les gros moteurs de recherche généraliste (type google) sont confrontés à cette « anonymisation » de l'utilisateur, les sites (Internet et intranet) des entreprises ne sont pas tout à fait dans le même cas.

En effet, les sites Internet des entreprises sont le plus souvent des « vitrines commerciales », or les services marketing des organisations sont là pour définir la clientèle potentielle de ces sites. Par conséquent, il leur est tout à fait possible de proposer des outils de recherche adaptés, sachant que les « critères » de recherche de cette clientèle sont souvent généraux et redondants (tarifs, moyen d'accès, horaires d'ouverture, catalogue des produits, mais aussi bilan annuel ou structure de l'entreprise).

En revanche avec les sites intranet, la situation est beaucoup plus complexe. Ici, le but n'est plus de fournir quelques informations « pratiques » à trois ou quatre profils types d'utilisateur, mais au contraire, de fournir des informations précises, techniques et régulièrement mises à jour, à des professionnels représentant tous les corps de métiers de l'entreprise. Or, si l'entreprise connaît « bien » les fonctions et les corps de métiers qui cohabitent en son sein, elle ne maîtrise pas toujours toutes les

subtilités de leur vocabulaire et de leurs processus métiers. Par conséquent, proposer des outils de recherche communs devient un véritable « casse-tête chinois » pour les concepteurs du portail.

Pour reprendre les propos de Sophie Azar – Exbrayat, la question qui se pose alors est : « C*omment faire en sorte que l'usager soit au centre du dispositif d'information ?* » [42 – AZAR-EXBRAYAT]

2.1.1 Comment fournir une information « ciblée » en entreprise ?

Ou comment transférer aux logiciels, les fonctions de reformulation des requêtes, de sélection des documents pertinents, et de classement des résultats, exercées autrefois par le documentaliste ?

Deux axes sont à prendre en compte :

2.1.1.1 Le contenu (les documents à organiser)

« Une inflation galopante des nouveaux textes [qui] rendront de plus en plus évident le besoin d'outils pour guider la recherche et décanter les données » [9-MANIEZ, p.165]

L'un des problèmes récurrents dans les grandes entreprises tient à la coexistence de plusieurs systèmes de gestions de contenus. Cette situation est le plus souvent la résultante de campagnes d'informatisation successives, parfois menées dans l'urgence afin de rattraper un retard technologique handicapant pour la survie de l'entreprise. Si, on se réfère à l'exemple d'Air France[30], l'ensemble de la compagnie utilise le système Lotus (IBM) qui intègre un service de messagerie, des espaces de travail collaboratifs, des modules de création de bases de données et des processus de workflow. Les fonctionnalités proposées par cette version de Lotus (4.6 et 5) ne permettant pas de créer des bases de données de type GED, plusieurs services ont du avoir recours à d'autres logiciels pour gérer et stocker leurs documents numériques

(ou fraîchement numérisés). Malheureusement, lorsqu'il s'agit de suivre l'évolution technologique des systèmes d'information, les entreprises ont souvent un temps de réaction proportionnel à leur taille. Par conséquent, elles se retrouvent fréquemment équipée de logiciels déjà dépassés au moment de leur implantation dans les services. Ainsi, il semblerait que l'outil retenu (Retriewalweare), pour faire office de logiciel de GED ait été déjà largement dépassé au moment de sa mise en application[31]. Par conséquent, au lieu de disposer d'un outil permettant de respecter l'intégrité des documents numérisés en les présentant sous un format PDF, les utilisateurs sont contraints de les consulter pages par pages dans un format RTF (rendu obligatoire par l'oscérisation), qui ne respecte pas la mise en page et les couleurs des textes originaux.
Un nouveau système implanté depuis peu, se présente comme un concurrent de taille en ce qui concerne l'accès aux documents : le portail Intralignes et son moteur de recherche Verity. Ce système de gestion de contenu, propose aussi de nombreux documents HTML créés spécialement pour les sites. Il s'agit pour la plupart d'articles issus des journaux électroniques consultables sur les portails métiers.

On constate aisément à quel point ces situations engendrent un foisonnement de fichiers électroniques de nature extrêmement variée : fichiers en format Word, Excell, PDF, RTF, format propriétaire Lotus, HTML, etc…Or, lors d'un processus d'indexation automatique, le format d'un document peut jouer un rôle capital. En effet, les documents « numérités » (cf.p.47) peuvent bénéficier d'une description de leur contenu à l'aide de balises intégrées dans la structure du document, c'est le cas des fichiers HTML, qui comportent des « métadonnées » invisibles pour le lecteur, mais lisibles par les moteurs de recherche. Mais, encore faut-il que les utilisateurs soient sensibilisés et formés à l'utilisation de ce langage…..

Néanmoins, l'avenir de l'indexation sur les portails web passe par ces techniques de représentation du contenu des documents. Celles qui offrent aujourd'hui le plus d'avenir, sont basées sur le langage XML (eXtensible Markup Langage ou Langage Extensible de Balisage).

---

[30] Voir p.104 Partie 3.1 : « Air France, une compagnie en quête de transversalité »
[31] Nous évoquerons ce point plus en détail dans la partie 3

« Comme HTML (Hypertext Markup Language) c'est un langage de balisage (markup), c'est-à-dire un langage qui présente de l'information encadrée par des balises. Mais contrairement à HTML, qui présente un jeu limité de balises orientées présentation (titre, paragraphe, image, lien hypertexte, etc.), XML est un métalangage, qui va permettre d'inventer à volonté de nouvelles balises pour isoler toutes les informations élémentaires (titre d'ouvrage, prix d'article, numéro de sécurité sociale, référence de pièce…) ou agrégats d'informations élémentaires, que peut contenir une page Web ». [74*]

Ce système de balisage est à la base des langages RDF (Ressource Description Framework) qui exprime des faits à l'aide de triplet d'URI (Uniform Resource Identifiers), et OWL (Ontology Web Language) qui fournit une syntaxe pour exprimer des relations logiques de type : union, intersection, inverse de, etc…
La combinaison de ces deux langages constituent les bases du web sémantique, et contribuent à aider les ordinateurs à mieux comprendre le sens des informations qu'ils traitent. [47-BORDAGE]

En résumé, le langage XML permet de structurer le contenu sémantique des documents dès leur conception. Toutefois, l'utilisation systématique de cette technique n'est pas encore à l'ordre du jour dans les grandes entreprises.

En effet, si ce mode de création des documents est envisageable pour le personnel travaillant fréquemment sur des documents conçus pour être publiés sur le réseau intranet, il n'en est pas de même pour les autres salariés. En effet, dans une organisation aussi tentaculaire qu'Air France, très peu d'employés ont réellement conscience que les capacités techniques d'un moteur de recherche comme Verity, permettent de visualiser par l'intermédiaire d'un portail web, les documents qu'ils produisent au sein de leur service. Si ce phénomène est entre autres, lié au fait que tous les services ne bénéficient pas d'un portail Intranet, il est aussi révélateur d'un manque de formation aux nouvelles technologies qui ne leur permettent pas d'imaginer l'existence de telles options. En conséquence, quels que soient leurs statuts ou leurs fonctions, la plupart des salariés continuent à produire des documents (le plus souvent au format Word) sans chercher à les « décrire », de manière à en

faciliter l'éventuelle exploitation sur le réseau Intranet.

C'est pourquoi face à un corpus où la nature électronique des documents est extrêmement variée, le recours à des « métamoteurs », capable d'effectuer des recherches en texte intégral sur des « containers[32] » hétérogènes (disposant ou pas de moteurs de recherche intégrés) est une solution parfaitement adaptée aux entreprises.

Toutefois, en l'absence d'un balisage contenu dans le corps du document, cette technologie ne garantie pas la pertinence de l'ensemble des documents retrouvés. Par conséquent, il est nécessaire d'envisager des solutions complémentaires pour guider en amont ou en aval la sélection des fichiers, de manière à ce qu'ils coïncident au mieux avec la requête des utilisateurs. Or, une entreprise est une organisation complexe, une sorte de modèle réduit de notre société, où cohabite des dizaines (dans le cas d'Air France, des centaines) de métiers différents, qui sont autant d' « univers » professionnels possédant leur propre vision de ce qu'est une information « utile ».

2.1.1.2 Les utilisateurs et l'activité de l'entreprise

« La difficulté, pour une unité d'information, d'appréhender l'adéquation entre l'information qu'elle collecte et organise, et la satisfaction de ses clients, est qu'une bonne part d'aléatoire entre dans l'appréciation de la situation de ceux-ci. Nous avons vu que l'information n'est pas but un en-soi, mais qu'elle dépend de la personne qui en a besoin et de l'exploitation qu'elle en fera dans le cadre de son travail. Il faut donc en permanence tenir compte de la situation et de l'activité des personnes susceptibles d'avoir besoin d'information déjà collectée ou à collecter. » [79-GUYOT, p.12]

[32] Terme employé par l'Editeur de Verity pour désigner les différents structures susceptibles de stocker des documents : base de données, dossiers électroniques, sites web….

L'information est devenue un élément décisionnel incontournable pour le personnel des entreprises. Mais, comment concevoir des systèmes de recherches d'information (et par conséquent d'indexation) adaptés à des cultures professionnelles différentes, cohabitant dans une même structure ? En effet, un commercial et un technicien appartenant à la même entreprise n'ont pas besoin de la même information. Par ailleurs, la fréquence et la forme sous laquelle est fournie cette information est conditionnée par le rythme et le cadre de leur activité professionnelle.

Il convient, avant tout, de réaliser un « référentiel culturel » des métiers de l'entreprise ou du service concerné. Pour mener à bien cette démarche, Eric Sutter propose d'envisager la question selon trois angles principaux [86-SUTTER, p.88]

- La nature des activités des professionnels de l'entreprise en terme de finalité.

  C'est à dire comprendre les objectifs et les missions confiés à chaque corps de métier, afin d'identifier les informations qui leurs seront utiles.

- La nature des besoins et des usages de l'information demandée.

  Savoir comment et dans quel cadre l'information sera utilisée permet de circonscrire le domaine de recherche :

  - informations internes : conditions tarifaires, état des stocks, etc.…
  - informations externes : étude de marché, veille concurrentielle

- Les « us et coutumes » de chaque corps de métiers

  Chaque profession a ses particularités, qu'il s'agisse des conditions de travail (en équipe ou seul), du lieu (fixe ou itinérant), du rythme (de nuit, de jour ou décalé). Tous ces éléments ont des répercussions sur le temps que les employés peuvent consacrer à la recherche d'information ou même simplement à sa prise de connaissance. Il est donc nécessaire pour le documentaliste de bien étudier les habitudes de chaque profession afin d'adapter au mieux la forme sous laquelle il transmet l'information aux utilisateurs.

La réalisation de ce « référentiel » peut prendre plusieurs formes. Le sujet de ce mémoire n'étant pas axé sur ce point, je me contenterai de les évoquer rapidement, sans développer plus avant les techniques nécessaires à leur mise en œuvre.

De même qu'on procède à une enquête de « besoins » ou « d'usages » sur l'ensemble d'un service pour définir les stratégies documentaires utiles, le documentaliste peut recourir à ce type de démarche pour analyser les méthodes de travail et les processus de recherche d'un groupe de salariés exerçant le même métier (ou les mêmes fonctions). Cette enquête prend en général la forme d'un questionnaire ou d'une série d'entretiens semi - directifs menés avec les salariés Toutefois, cette méthode a l'inconvénient d'être longue et coûteuse à mettre en place, ce qui la rend statique et difficile à mettre à jour.

Dans d'autres cas, le salarié peut configurer son profil en remplissant lui-même un formulaire. Cette méthode est plus souple mais elle ne donne qu'une vision interne de la démarche de l'utilisateur. Ce dernier manquant parfois de recul par rapport à leur propre fonction.

En dernier lieu, les spécialistes de l'information peuvent avoir recours à des profils générés dynamiquement. Un sous- système de modélisation observe l'utilisateur de derrière l'interface et découvre son profil à partir de ses actions. Cette méthode permet d'observer le cheminement de sa recherche sur le web (sur les sites Internet et sur le portail de l'entreprise), et elle corrige les caractéristiques de son profil au fur et à mesure de leurs évolutions. Ces dernières sont enregistrées dans une liste de variables, qui sont ensuite réutilisées par les logiciels, pour personnaliser les thématiques de classement des documents.

2.1.2 Personnaliser l'accès aux sources d'informations pertinentes

« La pertinence d'une réponse concerne d'une part l'exactitude de la réponse par rapport à la requête de l'utilisateur et d'autre part l'adéquation de la réponse par rapport au niveau de connaissance et aux préférences de l'utilisateur » [DAVID, p.341]

Les systèmes de recherche ne cessent d'améliorer leurs résultats en ce qui concerne l'exactitude des réponses, en proposant d'une part des variantes du langage booléen (logique modale, flou, recherche vectorielle), et d'autre part des « heuristiques pré-définies » basées sur des thésaurus pour mieux contextualiser la requête des utilisateurs. En rapprochant la question posée d'une ou plusieurs requêtes « type » [33], intégrées au préalable dans le système de recherche, il est alors possible de limiter les phénomènes de bruit et de silence au moment du processus de recherche.

En revanche, en ce qui concerne le second point : *« l'adéquation de la réponse ... »,* nous l'avons vu précédemment de nombreux paramètres liés au métier et aux fonctions exercés par l'utilisateur, rentrent en ligne de compte.

Aujourd'hui de plus en plus de logiciels intègrent des modules de DSI[34] (Alexandrie par exemple) qui établissent un profil de l'utilisateur en combinant les dernières références consultées par le salarié, avec la liste de ses centres d'intérêt préalablement saisis dans une fiche descriptive. Toutefois, ces techniques imposent à nouveau un exercice intellectuel qui consiste à traduire en « mots-clés » des termes et des concepts issus du langage naturel. Le manque de recul de l'utilisateur face à l'exercice de son métier et à son univers de travail rend souvent cette « auto-analyse » difficile et « partiale ».

Toutefois, d'autres solutions sont envisageables, l'une d'entre-elles consiste à demander aux salariés d'établir des « glossaires » métiers, permettant de clarifier l'usage et la signification des termes professionnels couramment utilisés dans leur corps de métier ou service. Le système de gestion de contenu, en s'appuyant sur l'annuaire d'entreprise est alors capable d'associer à chaque salarié (via son login), le glossaire correspondant. De cette manière, il peut affiner son processus de recoupement, entre les termes de la requête formulée par cet utilisateur et l'index issu de l'analyse automatique de l'ensemble du corpus de l'entreprise.

Cette technique est un premier pas vers une autonomie encore plus grande des utilisateurs. Toutefois, si on tient compte du volume des corpus stockés sur les

[33] Ces requêtes « type » sont utilisées par le moteur Verity sous le nom de Topic [33-Ivanciuc Deniau]
[34] Diffusion Sélective de l'Information

réseaux des grandes entreprises, la quantité de résultats obtenus est encore souvent trop importante pour ne pas dérouter les salariés. C'est pourquoi, il est devenu indispensable de proposer des solutions complémentaires, et plus pédagogiques pour guider les utilisateurs vers les bonnes sources d'informations.

2.1.3 Accélérer et faciliter l'accès aux documents

« *Comment, en amont, mettre à profit les savoir-faire documentaires pour permettre à l'usager, en aval, d'être autonome dans sa recherche d'information ?* » [42-AZAR-EXBRAYAT, p.197)

De nos jours, ce sujet mobilise tous les spécialistes de l'information. En effet, l'observation des processus de recherche sur Internet a permis d'établir que les utilisateurs « lambdas » allaient en général au plus simple, aussi bien dans la construction de leur requête[35], que lors de la consultation des résultats. Ainsi, on constate qu'ils se contentent le plus souvent d'inscrire un seul mot, au mieux deux ou trois, combinés essentiellement à l'aide de l'opérateur ET (plus rarement OU, et quasiment jamais SAUF). Par conséquent, le manque de précision de la requête, combiné à un mode de fonctionnement du moteur essentiellement statistique, induit un nombre extrêmement élevé de résultats. De plus, le classement des résultats est uniquement établi en fonction de la fréquence et de l'emplacement des mots recherchés et ne tient pas compte du contexte sémantique. Malheureusement, la plupart des utilisateurs ne le savent pas et ont pour habitude de consulter uniquement les 10 ou 15 premiers résultats. Par conséquent, s'ils ne trouvent pas la réponse à leur question sur la première page affichée, ils ne vont pas plus loin et en concluent amer : « On trouve tout et rien sur Internet, mais jamais ce que l'on veut !! ».

Cette situation se retrouve aussi fréquemment sur les réseaux Intranet des sociétés. Disposant de peu de temps pour effectuer eux-mêmes leurs recherches, les salariés des entreprises attendent de leur portail Intranet qu'il leur fournisse des réponses

[35] Observations recueillies lors d'une enquête menée par Altavista en 2001

précises en un temps optimal, puisqu'ils savent que cette information a été publiée sur le réseau[36].

Par conséquent, les documentalistes d'entreprise doivent mobiliser leurs efforts, sur la mise en place d'outils permettant de trier les informations en amont ou en aval de la recherche afin de délimiter le périmètre de recherche du moteur à un domaine correspondant à l'attente de l'utilisateur. En procédant ainsi, le spécialiste de l'information, contextualise le ou les termes recherchés, ce qui diminue les problèmes de polysémie.

Ces systèmes de filtrage des informations peuvent prendre plusieurs formes et intervenir à différents moments de la recherche.

En amont :

Ils se présentent, soit sous la forme d'une classification « active »[37] , qui permet de guider l'utilisateur au travers de thématiques de plus en plus précises, vers le domaine qu'il désire explorer. De là deux possibilités s'offrent à lui :

- Inscrire un terme dans un champ de recherche classique, où le moteur limitera sa zone d'exploration à la thématique choisie.
- Accéder directement à une série de documents, préclassés[38] dans cette thématique par la documentaliste, suite à une indexation manuelle, automatique ou semi-automatique.

Moins visible, mais néanmoins efficace, l'autre solution consiste à créer des requêtes « types », correspondant aux besoins les plus courant de chaque corps de métiers. A ces requêtes sont associées quelques documents pertinents, auxquels correspondent des combinaisons de « mots-clés », qui servent de références au moteur de recherche. Il s'agit en quelque sorte d'un enseignement par l'exemple, qui permet au

[36] Ou rendue accessible via un abonnement à des bases de données consultables par Internet.
[37] Qui dévoile sont arborescence au fur et à mesure de l'exploration, en cliquant sur les items, à la manière d'un lien hypertexte.
[38] On parle alors de classification « à priori »

moteur d'établir un système de correspondances sémantiques entre les différents documents du corpus d'une part, puis entre eux et les termes contenus dans les requêtes des utilisateurs. Si cette technique reste enfouie dans la mécanique interne des systèmes de gestion de contenu, lorsqu'elle est utilisée en amont de la requête, elle trouve une forme d'application très pédagogique en aval : Le topic Map.

En aval :

Il ne s'agit plus ici de guider l'utilisateur vers des « zones » de recherche correspondant à ses pôles d'intérêts, mais plutôt de trier et de classer les résultats obtenus, soit en fonction du profil de l'utilisateur, soit en fonction des thématiques intrinsèques au document.

- Dans le premier cas le moteur recensera tous les documents comportant des termes correspondant à ceux utilisés dans la requête (plus ou moins finement, suivant les traitements linguistiques dont dispose le logiciel), et les « dispatchera »[39] ensuite selon un plan de classement correspondant :
  - Soit au profil de l'utilisateur (établi grâce à un questionnaire personnalisé ou à partir d'un référentiel métier).
  - Soit à une classification fonctionnelle[40] de l'entreprise (taxonomies).
- L'autre option consiste, après le recensement, à proposer une « carte des concepts » (topic map) qui permet de présenter l'ensemble des résultats de la recherche sous la forme d'un « graphique » où les documents sont regroupés par lot. Chaque lot correspondant à la signification des termes en fonction de leurs contextes d'utilisation. Chaque domaine sémantique étant associé (ou pas) aux autres par des liens plus ou moins denses, en fonction de leur degré d'appariement.

[39] Principe du « clustering » (Katell Collet / URFIST Bretagne Loire-Atlantique 21/03/03 Comm LD Renadoc AvenirLD.doc p.24)

[40] C'est-à-dire qui ne respecte pas forcément l'organisation de l'entreprise telle que l'on peut la percevoir dans l'organigramme hiérarchique.

Ces classifications « a posteriori » résultent d'une indexation et d'un tri automatique, mais une intervention de « vérification/validation » humaine est le plus souvent nécessaire pour affiner les processus et supprimer certaines erreurs.

On constate ici, que ces outils, loin d'être des « usines à gaz », font tout simplement appel au bon sens et à la logique naturelle d'un processus de recherche humain. En réintroduisant des outils de classification avant ou après la recherche automatisée, les documentalistes reviennent aux sources de leur métier : guider les utilisateurs et leur fournir une réponse en adéquation avec leurs besoins.

Toutefois, ces fonctionnalités n'ayant pas été réintroduites dans les systèmes de gestion de contenu, par les documentalistes eux-mêmes mais par des informaticiens et des linguistes, leur dénomination a fortement évoluée. En quelques années, à la place des mots : classifications, plan de classement, liste vedettes matières, index, nous avons vu émerger dans la littérature spécialisée et dans les plaquettes de démonstration des éditeurs, des termes quelque peu obscurs tels que : taxonomies, ontologies, topics …..Seul le thesaurus semble avoir échappé à cette révolution lexicale, mais ces fonctions sont-elles toujours les mêmes ?

Il semble aujourd'hui nécessaire d'établir un « état de l'art » de la littérature professionnelle concernant ces nouveaux outils, afin de clarifier leurs fonctions et leur nature exacte. C'est ce que je vais tenter de vous proposer dans la suite de ce mémoire.

2.2 Les « nouveaux » langages classificatoires : des applications en constante évolution

« Anything that helps decrease the time it takes to employees or customers to perform a task or make a decision is a good investment » [49-PLOSKER,p.39]

C'est sans aucun doute ce « leitmotiv » qui mobilise depuis des années les concepteurs de logiciel et leurs principaux clients : les chefs d'entreprises. Des études menées en Californie et dans le reste du monde (altavista 2001) ont démontré que les salariés passaient environ 16% de leur temps de travail hebdomadaire à rechercher de l'information (soit pour un salarié français 5h30), 76% de ce temps étant uniquement consacré à la navigation. On imagine facilement quel est l'enjeu en terme de productivité, mais aussi du point de vue financier pour les entreprises. [32-SERRES , p. 3]

C'est pourquoi, depuis quelques années les éditeurs de systèmes de gestions de contenu ont repris à leur compte les recherches menées depuis un demi-siècle sur l'intelligence artificielle, en vue de créer des applications destinées à répondre à la question que se posent de plus en plus de spécialistes de l'information :

« Comment filtrer, comment réduire le nombre de références, comment exploiter les listes de résultats de manière plus « intelligente ». [….], bref comment mieux exploiter et gérer les informations. ». [32-SERRES]

Les recherches sur l'Intelligence artificielle visant à transférer aux machines, les principes du raisonnement humain, la première étape passait obligatoirement par une modélisation du langage naturel. C'est dans ce cadre que se sont développés les outils qui nous intéressent, leur point commun étant de s'appuyer sur des réseaux sémantiques.

Définition du réseau sémantique :

« *Un réseau sémantique est un graphe formé de noeuds qui représentent des concepts. Les arcs sont orientés et étiquetés. Il peut être structuré en plusieurs niveaux (domaine, concept, terme).* » (C.Leloup, Catégorisation et classifications automatiques, Journée ADBS, 11 avril 2002)

On constate aujourd'hui que les utilisateurs sont avant tout demandeur d'un système de guidage qui leur permet de se repérer dans le contenu informationnel de leur entreprise. Ils ont besoin de créer des liens, des ponts entre leur domaine d'activité et ceux de leurs collègues afin de mieux maîtriser la portée et les implications fonctionnelles de leurs missions. C'est pourquoi, de plus en plus « *l'outil* [de recherche] *adéquat devient à la place de la liste d'autorité, une grille de repères sémantiques échelonnés. On passe de l'univers de la linguistique formelle à celui de l'organisation du monde des connaissances »*

Ces nouveaux outils permettent l'utilisation de liens hypertextuels[41] qui abolissent la rigidité mono hiérarchique des classifications traditionnelles, au profit d'une polyhiérarchie transparente. Ce type de structure, plus flexible, est mieux adapté à la « variabilité » des besoins de l'entreprise.

Pour marquer ce changement de cap[42], les éditeurs, décidés à dépoussiérer l'image des ces langages classificatoires, ont délibérément choisi de les désigner à l'aide de termes n'évoquant pas (au premier abord) les sciences et techniques documentaires. Par la même occasion, ils se sont affranchis des normes qui contrôlaient les processus de création de ces langages, s'offrant ainsi une liberté de formes et d'utilisation, jusqu'à présent inédite. Cette situation nous amène à nous interroger sur les analogies et les différences existant entre ces deux générations d'outils.

[41] Le concept de l'hypertexte « reprend les principes d'organisation non - linéaire des contenus informationnels hébergés sur des machines distantes et accessibles par de multiples chemins de navigation. [….], les hypertextes et les hypermédias ont vocation à présenter l'information. » [thèse INPG, Martine Villanova-Oliver, 2002, p.1)

[42] D'aprèsA.Gilchrist, la consultation de la base LISANET, révèle que la première apparition du terme taxonomie dans son sens actuel date de 1997.

2.2.1 Les taxonomies

« Bien moins encore que les philosophes (…) les taxinomistes de la documentation ne sont encore parvenus à une théorie claire et cohérente des relations qu'ils manient, sans trop savoir apparemment de quoi il s'agit en fait. »

(Eric Grolier in « le système des sciences et l'évolution du savoir. Les fondements de la classification des savoirs. Munich : Verlag dokumentation, 20-119) [9-MANIEZ, p.212]

Il est vrai que si ce terme est devenu familier aux oreilles des spécialistes de l'information, sa nature exacte et surtout ses différentes applications, le rende difficile à cerner. Ce qui induit de nombreuses confusions. Bien souvent, il est utilisé en lieu et place des termes classifications, thésaurus ou même ontologies. On comprend mieux le trouble des documentalistes qui, en ce qui concerne les deux premiers outils, ont une idée très précise de ce à quoi ils correspondent.

L'analyse des articles publiés depuis cinq ans sur le sujet, tend à démontrer que cette confusion trouve son origine dans la nature même des taxinomies, qui leur permet d'être appliquées de manière si différente d'un système à l'autre, qu'elles en deviennent parfois, difficilement détectables.

C'est pourquoi, nous allons dans un premier temps, définir les origines de ce terme, avant d'étudier son degré de parenté avec les autres langages classificatoires.

2.2.1.1 Définitions

Origine du terme

Taxinomie : du grec *taxis* « mise en ordre » et *nomos* « loi »

Un débat agite les spécialistes depuis la reprise de ce terme par les spécialistes de l'information. Doit-on dire Taxinomies ou Taxonomies ?
On trouve de nombreux articles sur la question sur le site de l'Encyclopédie communautaire du web : Wikipedia. Toutefois, c'est l'article publié par *Le grand*$_{80}$

dictionnaire terminologique de l'Office Québécois de la langue française qui m'est apparu comme le plus complet. Il indique que le terme « taxinomie » est plus approprié que « taxonomie » car ce dernier serait un calque du terme anglais « *taxonomy* ». Pourtant dans les dictionnaires Anglais, ce dernier est désigné comme étant d'origine française et venant du terme « taxonomie », crée en 1813 par un botaniste suisse (A.P. Candolle). [50-SCHEIDER]

Plus simplement, cette variation tient sans doute au fait que, dans le domaine scientifique et dans la plupart des langues, le « o » prévaut le plus souvent sur le « i » pour établir une liaison entre deux termes[43]. Mais, peu à peu, on a pu remarquer une répartition de l'usage de ces deux termes : D'un coté, les biologistes et autres scientifiques optant pour taxonomie (n'oublions pas que le terme a été crée par un botaniste) et de l'autre les linguistes et autres spécialistes de la documentation, préférant utiliser le terme « taxinomie ».

Jusque là, la situation était encore à peu près claire, mais tout se complique, à partir du moment où les éditeurs de logiciels se sont emparés de l'expression. Produits par des sociétés à vocation internationale, ces outils (et leurs vendeurs) ont très vite opté pour le terme anglais « taxonomy », en lieu et place de nos « taxinomies ». Au moment de l'implantation chez le client français, la *taxonomy* devient alors une taxonomie et le salarié chargé de les concevoir, un(e) taxonomiste.

Nous l'avons vu l'origine de ce terme est associée au monde de la biologie et aux premières classifications des espèces vivantes (Linné, fin XVIIIème siècle). Leur principe d'organisation qui permet d'aller du général au particulier, est celui que l'on retrouve dans l'usage actuel des taxinomies. Ces derniers permettent de décrire un domaine (ou le sujet d'un corpus de document), en organisant les différents concepts qu'il contient dans une classification hiérarchique, organisée selon un point de vue unique (un seul type de relation entre les termes pour chaque taxinomie).

[43] Dans la plupart des langues d'origine latine le « o » a prévalu sur le « i » pour construire les termes : Tassonomia (it) , taxonomia (pt), et même le danois propose « taxsonomi ». Mais, en Grèce et en France, on dit taxinomia et taxonomie.

Si cette relation de spécialisation[44] est la plus courante, on trouve aussi d'autres systèmes d'organisation reposant sur des relations d'instanciation (les termes sont reliés par « est un ») ou de partition (lien du type « est une partie de »).

Cette variété de relations, permet de créer des classifications « sur mesure » en fonction de leur domaine d'utilisation. Intégrées dans les systèmes de gestion de contenu, elles ont absorbé les technologies hypertextes, ce qui va alimenter un certain glissement de leur « fonction » dans l'esprit des spécialistes de l'information. Instruments conçus à l'origine pour classer les informations, les taxinomies sont devenues des outils de présentation et d'accès aux données.

Dans un excellent article publié en 2003, par le « Journal of documentation », Alan Gilchrist recense dans la littérature spécialisée, au moins cinq applications différentes, désignées sous le nom de taxinomies. [58 - GILCHRIST]

2.2.1.2 Cinq domaines de prédilections pour les taxinomies

2.2.1.2.1 Les répertoires Web (web directories)

Très présent sur les sites Internet (yahoo et altavista), ils commencent aussi à s'imposer sur les réseaux d'entreprise. Ils sont particulièrement employés par les portails où ils sont plus connus sous le nom d'annuaires. Ils proposent aux utilisateurs un « menu » composé de « *top terms* » qui permet de lister les principaux domaines abordés dans le site. En cliquant, sur l'un ou l'autre de ces termes, on accède à des séries de concepts illustrant, de plus en plus finement, les différentes facettes du domaine exploré. Pour mieux rendre compte de la diversité « culturelle » des utilisateurs, certains termes peuvent être répétés à plusieurs reprises et à différents niveaux de l'arborescence. Ainsi, on multiplie les portes d'entrées vers l'information recherchée.

Lorsque l'utilisateur arrête son choix sur le concept qui lui convient, il peut accéder directement aux documents qui ont été classés dans cette catégorie ou entrer le terme choisi dans un moteur de recherche. Ces deux types d'accès sont proposés par les taxinomies mises en place à Air France.

[44] Chaque terme de l'arborescence est relié aux autres par une relation signifiant « est une sorte de ».

2.2.1.2.2 Des taxonomies pour guider (trame) l'indexation automatique

Alan Gilchrist illustre cette application en évoquant le cas des sites commerciaux. Si, l'on observe un site comme laredoute.com, l'utilisateur peut naviguer dans une classification simple (pas plus de deux niveaux), qui permet d'accéder aux informations et aux photos illustrant les produits. Mais le consommateur peut aussi effectuer directement une recherche en inscrivant un mot dans le champ d'interrogation. Ce choix implique de gérer les problèmes de synonymie et de syntaxe qui en découlent. [58 – GILCHRIST]

Dans ce contexte, il devient nécessaire de dissimuler derrière chaque terme de la classification, une taxonomie qui regroupe l'ensemble des synonymes de ce mot, toutes les formes erronées d'écriture qui peuvent lui être associées et les règles de gestion qui s'imposent dans chaque cas. Dans ce contexte, les taxonomies s'apparentent à des thésaurus (puisqu'elles mettent en place un système de relation du type « employé pour ») qui viendrait enrichir les classifications mises en place sur le site. Par exemple, derrière la catégorie « articles de puériculture », on pourrait trouver la taxinomie suivante :

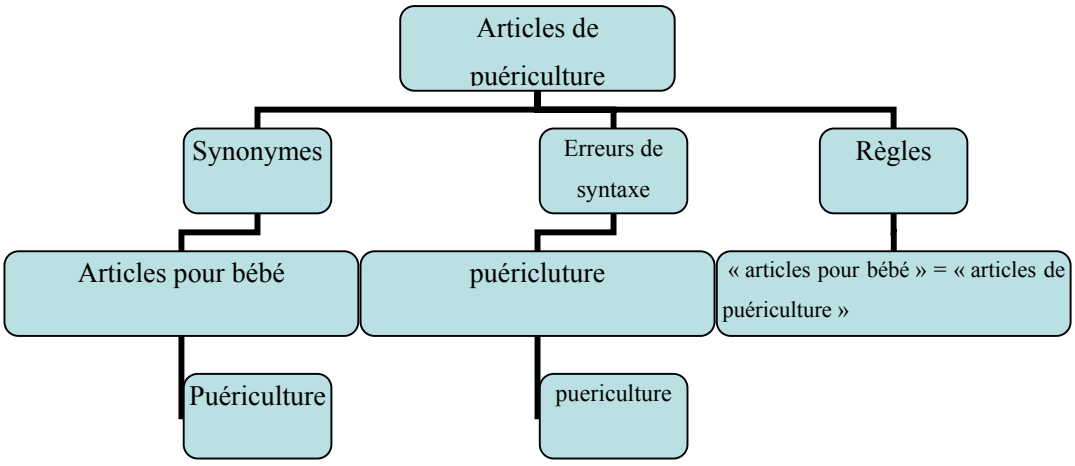


Figure 3 : exemple de taxinomie pour un site web commercial

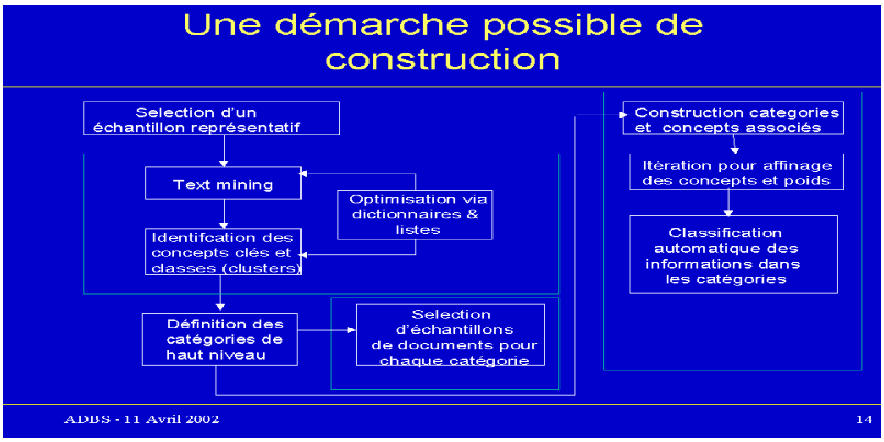
Ce principe permet d'affiner l'indexation automatique des gros corpus de documents qui serait impossible à traiter manuellement. En revanche, la classification et les taxonomies qui lui sont associées sont élaborées manuellement.

2.2.1.2.3 Des taxonomies pour créer des classifications automatiques

En combinant les taxonomies avec les outils d'analyse sémantique et linguistique, les moteurs de recherches, identifient les concepts présents dans les textes, les liens qui existent entre ses concepts et leur poids au sein du corpus. La combinaison de toutes ces données, permet d'établir une arborescence thématique, dans laquelle le moteur redistribue les documents analysés en fonction des différents concepts qu'ils contiennent. Les catégories ainsi créées peuvent réutiliser sur les portails intranet sous la forme de répertoires proposant un accès hypertextes aux documents stockés sur les serveurs de l'entreprise. Cette application est proposée en option dans le logiciel Verity, sous le nom de « Verity intelligent classifier ». Air France, en le combinant avec ses glossaires « métiers », va utiliser ce dispositif pour proposer à ses salariés un classement automatique des résultats de leurs recherches en fonction de leur profil.

Les « topic maps », que nous évoqueront en détail un peu plus loin, sont en quelques sortes la traduction bidimensionnelle de ces classifications automatiques.
Le principal avantage de cette technique est de permettre une mise à jour rapide des « sujets » proposés sur le réseau intranet, au fur et à mesure de la publication des nouveaux documents.

Figure 4 : Scénario pour construire une taxinomie C.Leloup- Journée ADBS-11/04/2002



2.2.1.2.4 Front and filters : la taxonomie comme outil de navigation et de pertinence

Dans cette optique, la taxonomie mise en place sur le portail, se substitue au moteur de recherche. Au lieu de formuler sa requête dans un champ d'interrogation, l'utilisateur se déplace dans l'arborescence pour sélectionner le contexte d'utilisation du terme qu'il recherche, ce qui diminue efficacement les problèmes d'ambiguïtés.

Dans le cas d'Air France, le terme Concorde désignant à la fois un journal interne et l'avion, la taxinomie permet de contextualiser ce terme soit dans la catégorie « communication interne », soit dans celle de « la flotte aérienne ». Cette démarche présente l'avantage de pouvoir découvrir visuellement les domaines en corrélation avec le thème de la recherche principale. Ainsi, l'utilisateur peu « dérouler » chaque thématique, en passant de l'une à l'autre, au gré des associations d'idées que cela lui inspire. Le thesaurus « dynamique » de la BDSP, est un exemple de ce type d'application.
Cette approche plus pédagogique de la recherche, offre aux utilisateurs une meilleure perception de leur entreprise, et une meilleure maîtrise des principes de la recherche documentaire. Ce qui lui permet d'enrichir la formulation de ses requêtes, lorsqu'il doit les formuler sous formes de mots-clés dans un champ d'interrogation.

2.2.1.2.5 Les taxonomies d'entreprise

« Ces instruments servent bien entendu au classement et à la représentation des informations ; mais ils sont conçus pour refléter et projeter une conception des métiers, des savoir-faire, des modes de fonctionnement de l'organisation, bref une culture de l'entreprise » [64 – MENON]

Depuis le milieu du XXème siècle, la politique économique mondiale incite les entreprises à se regrouper pour former de grands consortiums. Fusions, OPA et démantèlements, exigences du marché, nouvelles techniques de managements se succèdent, avec pour conséquences des restructurations incessantes des entreprises.

Au sein du service DP.SQ d'Air France, j'ai pu constater que l'assistante du chargé de communication (dont l'une des missions consiste à maintenir à jour la liste des

représentants du réseau), était confrontée presque quotidiennement, à la disparition et à l'apparition de nouveaux services. Seul problème, l'organigramme hiérarchique de la compagnie n'est remis à jour qu'une à deux fois par an. De plus, il ne propose que les premiers niveaux de son arborescence. Conclusion, le seul point de repère dont dispose les salariés est l'annuaire Lotus du personnel, malheureusement il ne semble pas « connecté » avec la base de données des ressources humaines, puisque là aussi on trouve des salariés partis à la retraite depuis quelques mois !! Certes, cette situation est anecdotique, cependant elle est révélatrice des disfonctionnements inhérents à ces grandes sociétés où des systèmes d'information hétérogènes sont amenés à cohabiter aux grés des fusions.

Structure tentaculaire, administration parfois kafkaïenne, dispersion géographique, équipements informatiques hétérogènes, réseaux informatiques parfois sous dimensionnés par rapport au volume des documents et aux nombres d'utilisateurs, sont autant de paramètres qui rendent difficile l'accès à l'information pour les salariés.

Or, ces informations sont devenues des outils stratégiques pour la survie des entreprises. Quelles soient légales, économiques ou techniques, elles font toutes l'objet d'une veille qui permet aux dirigeants d'orienter leur politique managériale. Avec l'explosion d'Internet et des bases de données spécialisées, la circulation des informations s'est tellement accélérée que l'enjeu de nos jours, ne consiste plus seulement à trouver la bonne information mais à la trouver avant les concurrents.

Cet impératif est également valable pour les informations internes, c'est pourquoi il est devenu nécessaire d'organiser le capital informationnel de l'entreprise. Mutualiser les applications informatiques, harmoniser les supports de publication, guider rapidement les salariés vers les sources d'informations qui correspondent à leurs besoins, ces objectifs sont devenus des priorités pour les manageurs des systèmes d'informations.

Combinées avec un métamoteur (type Verity ou Arisem), les diverses applications des taxinomies, peuvent fournir des solutions réellement efficaces aux problèmes que nous avons évoqués. Utilisées dans le cadre d'un portail intranet, ces taxinomies permettent de dresser une carte fonctionnelle de l'entreprise, que les salariés peuvent utiliser pour explorer les divers secteurs d'activité de leur organisation. Reprises sous forme de répertoires, elles permettent de structurer l'accès aux documents internes et externes, stockés dans les multiples bases de données des services. Par la même occasion, elles permettent d'affiner et de cadrer l'indexation automatique des documents en introduisant des « règles » pour gérer les problèmes de synonymie et de polysémie (à la manière d'un thésaurus).

La dernière tendance pour mieux décrire les liens fonctionnels ou hiérarchiques, existant entre les métiers et les services d'une entreprise, consiste à combiner les taxinomies avec des ontologies. Ces dernières permettent d'établir des liens sémantiques entre les termes du type « collabore avec », « participe à », etc…

En conclusion, on peut dire que le terme « taxinomie » est un terme générique, qui désigne une structure hiérarchique pouvant prendre plusieurs formes, suivant l'organisation choisie (facettes) ou les relations sémantiques qu'on lui adjoint. Par conséquent, quel que soit le langage classificatoire étudié (thésaurus, topic maps, ontologies), son noyau structurel est une taxinomie. Seul « l'emballage » change, c'est-à-dire le type de relation qui unit les termes entre eux, ses fonctions au sein du système de gestion de contenu et la forme sous laquelle il présente les documents aux utilisateurs.

2.2.2 Les Thesaurus

2.2.2.1 Définition

L'OED (Oxford English Dictionnary) définit les thesaurus comme: "*A classified list of terms, especially keywords, in a particular field, for use in indexing and information retrieval*". [58 – GILCHRIST]

Cette définition est très semblable à celle de l'AFNOR (cf. partie 1, p. 40) que nous avons évoquée dans la première partie de ce mémoire, consacrée aux usages traditionnels des langages classificatoires.

R.Texier, dans un article pour Elikya (concepteur de logiciel), désigne le thésaurus comme « *une taxonomie sur laquelle on ajoute des relations « horizontales » en plus de la relation verticale qui structure la taxonomie originale. Ces relations expriment une connexité entre sujets par des liens du type « relatif à », « voir aussi » ou « synonyme de », ...et permettent d'élargir le champ d'étude ou de recherche.* » [48-TEXIER]

Dans les deux cas, on observe que le fonctionnement interne de cet outil est le même, en revanche les fonctions qui lui sont attribuées ne sont pas tout à fait identiques.

2.2.2.2 Applications

Dans la définition de l'OED (et de l'AFNOR), le thésaurus est présenté avant tout comme une classification fournissant des « mot-clés » pour indexer des documents (et les retrouver), alors que la définition proposée par un éditeur de logiciels utilise immédiatement le terme de taxonomie à la place de classification, et semble le présenter comme un outil d'exploration. C'est dans cette nuance que se situe essentiellement l'évolution des thésaurus.

Transposé dans l'univers d'Internet, le thésaurus a hérité de ses fonctionnalités hypertextes. Ce qui permet de proposer une exploration dynamique des champs

sémantiques du thésaurus et de fournir un accès direct aux documents.

Dans ce contexte technologique, le principe des thésaurus à facettes a rencontré un succès particulier au près des informaticiens. Jacques Maniez définit les facettes comme :

« Des composantes essentielles d'une structure complexe, dont la séquence ordonnée peut réaliser une recomposition et un classement rationnel des entités soumises à cette structure. » [9-MANIEZ,p. 228]

Et l'AFNOR, comme :

« *Des catégories de notions de même nature ou exprimées d'un même point de vue telle que phénomène, processus, propriété, outil, permettant un regroupement des notions indépendamment des disciplines traitées. ».* [22-AFNOR]

Réutilisées par les informaticiens, ces « facettes » sont des structures permettant de regrouper les concepts qui correspondent à tous les « angles » d'approches envisageables pour une même rubrique. En dotant ce type de thésaurus, d'interfaces ergonomiques, ils deviennent alors d'excellents outils de navigation, « précis et flexibles », qui permettent une exploration allant « du plus large au plus étroit », tout en suggérant à l'utilisateur toutes les combinaisons susceptibles d'exister entre ces concepts. Le thésaurus en ligne du BDSP[45] illustre très bien les nombreux avantages liés à cette nouvelle utilisation des thésaurus. Il propose deux moyens d'accès aux documents, le premier consiste à se déplacer dans la représentation graphique du thésaurus, en exploitant les liens dynamiques reliant les rubriques entre elles, pour sélectionner le « concept » qui permettra d'accéder aux documents pertinents.
Le second est un moteur de recherche classique, où l'utilisateur inscrit le terme qu'il recherche dans un champ. Il peut être guidé dans sa démarche par une liste d'autorité qui reprend uniquement les descripteurs du thésaurus.

[45] banque de données de la santé Publique (www.bdsp.tm.fr)

Ces thésaurus de recherche[46] présentent encore une fois de grandes similitudes avec les *topics Maps*. Comme eux ils gèrent des réseaux de concepts, dont ils proposent une représentation graphique, dynamique et évolutive. La parenté est encore plus grande, lorsqu'ils sont enrichis par *« des ontologies des connaissances sur les personnes [les utilisateurs], les lieux ou les produits »*, qui leur octroient une plus grande précision dans le traitement automatique de l'indexation, puisqu'elles permettent de la personnaliser et de la contextualiser. [64 – MENON]

2.2.3 Les Topics Maps

Topic maps : le terme est utilisé pour décrire un standard ISO pour la représentation et l'échange de connaissances, en mettant l'accent sur l'accès à l'information. Le standard est formellement identifié sous la référence ISO/IEC 13250:2003.

« *Un topic map représente une information en utilisant des « sujets » (topics » en anglais) qui représentent tout concept, tel qu'une personne, un groupe de personnes, une couleur, un pays, une organisation, un module logiciel, un fichier individuel, des évènements, des « associations » qui représentent les relations entre ces « sujets », et des « occurrences » qui représentent des relations entre des sujets et des ressources informationnelles qui s'y rapportent. L'intérêt des topics maps est de définir des contextes et profils d'utilisateurs particuliers, et faciliter la fusion de topic maps provenant de sources différentes*. » [64*}

Dans une topic map, chaque sujet (topic) peut-être associé à un nombre illimité de synonymes, et un même terme peut être associé à plusieurs concepts (ce qui est interdit dans les thesaurus). L'assemblage de plusieurs « topic » permet de former un « scope », terme qui désigne le contexte d'utilisation le plus approprié pour le terme recherché.

Le principal atout de cet outil est de proposer une navigation plus pédagogique, et une vision d'ensemble du sujet principal de la recherche et des concepts qui peuvent lui être associés. En lui associant des ontologies, il devient capable d'établir des

[46] De « search thesaurus », employé par A.Gilchrist (journal of documentation, p.9) [58]

correspondances entre différentes classifications (taxinomies, index, thesaurus ….) et de les fondre dans une représentation graphique commune.

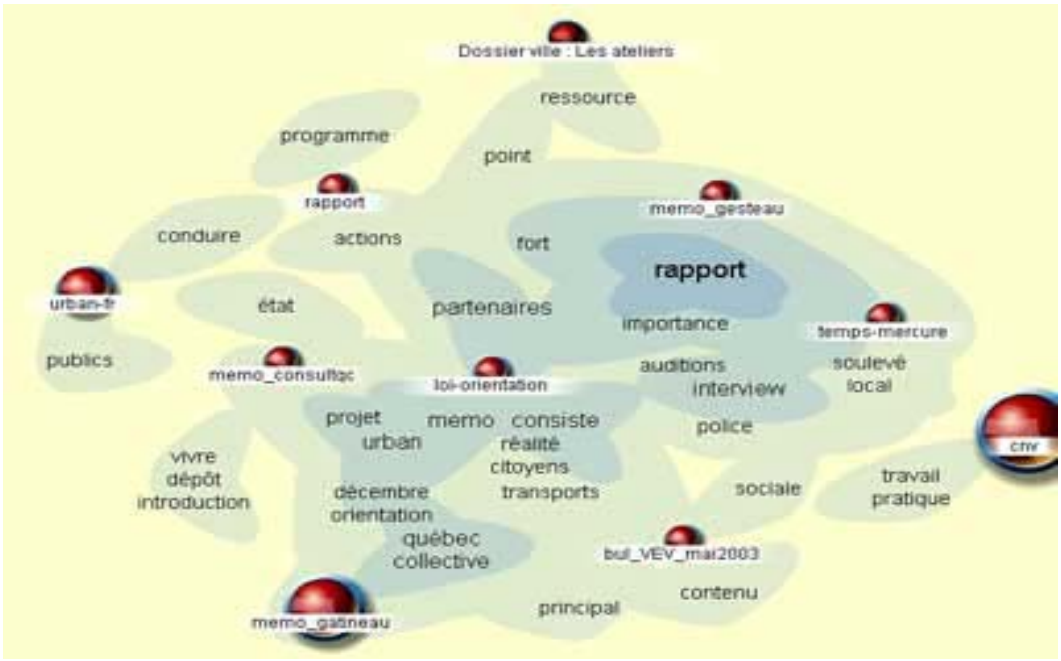


Figure 5 : Interface graphique d'un projet de topic maps pour le site www.planetecologie.org , produit KARTOO

Cette flexibilité permet aux topic maps de suivre l'augmentation rapide des corpus de documents au sein des entreprises. Ils permettent aussi d'harmoniser les accès à ces informations en intégrant le vocabulaire issu des différents outils d'indexation (thesaurus, taxonomies, classifications à facettes, listes d'autorités) qui coexistent au sein des entreprises en raison de l'hétérogénéité des applications mises en place dans les services.

2.2.3 Ontologies

« *Pour être susceptibles d'être exploitées automatiquement, les métadonnées doivent être entièrement explicites, c'est-à-dire suivre un modèle et être exprimées dans un vocabulaire clairement et formellement défini. Les ontologies, deuxième pilier du Web sémantique, sont le réceptacle de ces définitions. Elles modélisent les connaissances nécessaires à la description – et au traitement – d'un ensemble de ressources. On y représente les valeurs que l'on peut donner aux métadonnées et l'interprétation que les systèmes peuvent en faire, c'est-à-dire les concepts d'un domaine, les relations qu'ils entretiennent et la sémantique de ces relations, mais aussi les règles de raisonnement qui leurs sont applicables. » [65-GRILLL, p. 40]*

Cet extrait du livre Blanc du GRIIIL sur le traitement automatique des langues dans l'Industrie de l'information, insiste sur l'orientation technique des ontologies. Contrairement au thésaurus, qui sont conçus pour être des outils de médiation entre les sources d'informations et les humains, les ontologies sont destinées à faciliter les échanges entre les applications informatiques. En représentant les multiples aspects des ressources numériques à l'aide de relations sémantiques, elles potentialisent l'exploitation des langages documentaires (thésaurus ou taxonomies) par les logiciels de traitements linguistiques automatisés.

2.2.3.1 Définition

« An ontology can be defined as a normal, explicit specification of a shared conceptualisation » (GRUBER, what is an ontology? www.ksl.Stanford.edu/kst/what-is-an-ontology.html). [58-GILCHRIST, p. 13]

"*An ontology describes the subject matter using the notions of concepts, instances, relations, functions, and axioms. Concepts in the ontology are organized in taxonomies through which inheritance mechanisms can be applied"*(Benjamins, V.R et al (n.d),"(KA)2 : building ontologies for the internet : a mid term report www.cs.vu.nl/~dieter/ftp/paper/km99.pdf) [58-GILCHRIST, p. 13]

Les origines du terme ontologie appartiennent au domaine de la philosophie et plus particulièrement à la métaphysique et désignent la discipline qui explore la nature fondamentale du monde réel. Transposé dans le domaine des sciences documentaires, ce terme désigne des classifications (taxonomies) qui *« structurent les termes d'un domaine en établissant entre eux des relations de proximité, notamment des relations verticales de type « est un » ou « partie de ». Celles-ci permettent de mettre en mémoire, sous le nom d'« héritage », les attributs d'une classe d'objets que l'on retrouve dans ses espèces ou dans ses parties. »* [9-MANIEZ, p. 239]

2.2.3.2 Fonctionnalités

Ce « mécanisme » de l'héritage, est sans doute l'un des concepts les plus porteurs pour le développement futur des logiciels de gestion de contenu. En effet, il permet d'attribuer un certain nombre de propriétés aux principales catégories d'une classification et de les transmettre à tous les concepts qui lui seront assignés.

Katherine Adams dans son article « The semantic Web » , illustre cette propriété des ontologies, par l'exemple suivant :

Si on prend une classification consacrée aux chanteurs, et que l'on sélectionne l'artiste Ricky Martin dans la catégorie « Pop Star ». Une ontologie peut permettre l'établissement d'une règle qui dit que « si un artiste a un agent ou un manager, et qu'il a réalisé un album au cours de l'année précédente, il doit avoir un fan club, en conséquence le logiciel de recherche peut en déduire que Ricky Martin a un fan club, et lancer une recherche pour le trouver. [60-ADAMS]

Les ontologies en formalisant et en explicitant les relations entre les termes d'une classification, permettent d'effectuer des recherches portant justement sur les liens qui unissent ces catégories, par exemple dans une ontologie du type « X mange Y», on peut poser une question du type : « qui mange des souris ? », ou « que mange le chat ? »

Les ontologies sont la forme la plus « puissante » des langages permettant de décrire le sujet d'un document. Leur ascendant sur les autres outils, tient en particulier au fait qu'il s'agit d'un système complètement ouvert et évolutif, contrairement aux taxinomies et aux thésaurus.

En effet, si le nombre de catégories dans une taxinomie n'est pas limité, en revanche il ne peut exister qu'un seul type de relation entre elles (toutefois la nature de ces relations peut varier d'une taxinomie à l'autre). Dans le cadre d'un thésaurus, la situation est inversée, le nombre de relations est plus élevé, ce qui permet une meilleure indexation, mais le nombre de concepts disponibles est très difficile à augmenter.

« On peut voir dans les ontologies les descendantes surdouées des thésaurus, encore au berceau, mais susceptibles une fois parvenues à maturité de devenir des outils vraiment universels de caractérisation de l'information » [64-MENON, p.3]

L'analyse structurelle de ces nouveaux systèmes de gestion des connaissances, démontre qu'ils sont les héritiers des langages documentaires traditionnels. En synthétisant et en adaptant leurs fonctionnalités au monde du web, ces nouveaux outils ont dépassé les clivages qui opposaient les langages classificatoires aux langages combinatoires, les pré-coordonnés au post-coordonnés, etc…Créés pour être utilisés sans intermédiaire, par un public souvent néophyte, ils intègrent dès leur conception, des notions d'ergonomie et une grande souplesse d'adaptation. Ces nouveaux outils, loin de s'opposer ou de s'exclure mutuellement, ont des champs d'action complémentaires (voir Annexe 1). En les combinant dans un même système de gestion de contenu, ces technologies augmentent l'efficacité des moteurs de recherche et améliorent l'autonomie des utilisateurs. A la fois plus pédagogiques et plus intuitifs, ces outils permettent une rationalisation et une personnalisation de l'accès à l'information, gage d'une meilleure circulation de l'information dans l'entreprise.

2.2.3.3 Applications

Aujourd'hui le but des chercheurs en ingénierie documentaire, est de réussir à créer des ontologies universelles, qui pourraient être appliquées et réutilisées dans toutes les disciplines, afin de faciliter l'échange et l'exploitation de l'information numérique. Cette démarche est au cœur d'un des plus grands projets de l'Intelligence artificielle : le web sémantique. Ce dernier s'appuie sur le OWL (Web Ontology Language), un dialecte XML basé sur une syntaxe RDF, qui permet de décrire les catégories et les propriétés qui leurs sont attribuées. Grâce à cet outil, mis au point par W3C (World Wide Web Consortium), il sera envisageable de formuler de vraies requêtes en langage naturel.

Au sein de l'entreprise, les ontologies permettent de répertorier et de décrire les concepts métiers propres à l'entreprise, ce qui joue un rôle essentiel dans le cadre d'une politique de knowledge management. En organisant les informations disponibles sur le réseau, en fonction de ces mêmes concepts, les ontologies permettent une lecture uniforme de l'information dans toute l'entreprise.

2.3 Les logiciels leaders sur le marché

"Face au "déluge informationnel", la question n'est plus tant la recherche elle-même que le traitement, l'exploitation des résultats » [9-MANIEZ]

Depuis la démocratisation de l'accès à Internet, on observe un phénomène d'échanges technologiques entre le monde des logiciels documentaires et celui des outils du web.

D'une part, on observe une « **webisation** » des bases de données professionnelles. Ces dernières, en se dotant d'une interface web et des outils qui lui sont habituellement associés (moteur de recherche en texte intégral, navigation par liens hypertextes), deviennent plus accessibles aux « non spécialistes de l'information », qui sont de plus en plus familiarisés avec les principes de la navigation hypertexte.

Parallèlement, les moteurs de recherche sur le web (google, yahoo, etc ..) proposent de plus en plus souvent des « aides à la recherche » qui s'appuient sur les langages documentaires :

- Système de « guidage » via un thésaurus pour affiner les recherches, ou élargir le champ d'investigation.

- Classifications automatiques pour trier les résultats en fonction de leur pertinence selon le profil de l'utilisateur.

Par conséquent, on assiste à une profonde modification du paysage commercial des éditeurs de logiciel. Ainsi, en 2004, Google, le spécialiste de la recherche sur Internet, a lancé son premier système de gestion de contenu Intranet. Parallèlement, beaucoup de petits éditeurs incapables de lutter face aux leaders du marché, ont recentré leur activité sur la conception d'outils très spécialisés qui peuvent venir se greffer sur d'autres applications (intégration du module de traduction de lingway dans les applications de VerityK2). Cette explosion étant relativement récente, il existe peu de littérature « objective »[47] permettant d'analyser et de comparer les performances de ces outils. Réduite à la consultation des plaquettes commerciales des éditeurs, j'ai préféré limiter cette présentation aux applications qui connaissent le plus grand succès auprès des entreprises.

- **VERITY K2**

Leader incontesté du marché, quelque soit la technologie de base du moteur (à dominante statistique ou linguistique), ce moteur est produit par la société américaine Verity, implantée en Californie. Crée en 1998, elle emploie 580 salariés dans le monde et 11500 clients internationaux font appel à ses services dont 1200 répartis entre la France et le sud de l'Europe (Air France, Airbus, Alcatel, France Telecom, L'INIST, etc.). Elle génère un chiffre d'affaire de 124 millions de dollars, en constante augmentation de puis 2003 (102 M$). Ce système peut prendre en charge 250 formats et 26 langues. Mais il ne peut effectuer des opérations de cross-linguisme qu'en intégrant un module développé par son partenaire Lingway. [33 – INVANCIUC DENIAU]

[47] le dernier tableau comparatif sur les moteurs de recherche publié par le journal du net date de 2003.

Verity K2 se décline en 3 produits :

- **Verity K2 developper** : pour le développement des applications informatique (programmation)
- **Verity K2 Catalog** : pour les solutions du e-commerce
- **Verity K2 Entreprise** : pour la gestion des intranets d'entreprise

Ce dernier a été retenu par Air France pour gérer son réseau intranet et son portail. Sa technologie s'articule autour de 3 axes :

- L'information interne
- L'information externe
- Les processus de fonctionnement de l'entreprise

Son moteur en texte intégral incorpore des processus d'analyse sémantique, statistique et syntaxique qui permet d'effectuer des recherches aussi bien dans l'information structurée (base de données) que non structurées (fichiers bureautiques). Il permet de construire des requêtes booléennes complexes (35 opérateurs)

Il propose deux modes d'organisation de l'information :

Statique : à l'aide de taxinomies créées en collaboration avec le personnel de l'entreprise. Les documents y sont rangés manuellement ou automatiquement à l'aide des métadonnées, ou grâce à des lots de documents « référents » sur lesquels s'appuie le logiciel, pour étalonner son processus de classification. Cette dernière méthode suppose une vérification humaine régulière.

Dynamique : le moteur extrait, seul, les thématiques du ou des documents et les organise sous la forme d'une classification dans laquelle il présente les résultats de la recherche. La sélection des thématiques peut-être orientée en fonction du profil de l'utilisateur qui procède à la recherche.

Les taxinomies sont gérées dans **VIC** (Verity Intelligent Classifier), ce logiciel permet de créer les classifications, les requêtes pré-enregistrées et le ranking, il met

aussi à la disposition des collaborateurs un processus de workflow et des groupware.

- **AUTONOMY [27]**

Cette société se place au 1er rang des moteurs de recherche à dominante statistique, son offre tourne autour de 3 produits clés : IDOL Derver, Autonomy Server, Portal-in-a-Box). Son chiffre d'affaire en 2004, était de 64M$ en progression de 18%. Jusqu'à présent spécialisé dans le traitement des informations non structurées, il a racheté la société californienne Ncorp (éditeur de solutions logiciel pour le traitement de données structurées) afin de se placer sur le marché de l'e-commerce, en accédant au traitement des formulaires. Elle compte parmi ses clients BAE Systems, Ford, Sun Microsystems, IBM, la NASA, et le Département de la défense américain.

Le produit proposé grâce à cette fusion s'appelle : « Autonomy Directed Navigation ». Il propose une recherche avancée à l'aide d'un processus de navigation en temps réel dans des taxonomies multiples, y compris l'accès complet aux intersections de définitions taxonomiques diverses. Il n'exige aucune intervention manuelle supplémentaire.

- **CONVERA**

Anciennement Excalibur, puis Acamaya, cette société possède 800 clients répartis sur 33 pays.

Son produit Retriewalware 8 gère 200 formats de fichiers et 45 langues, sa structure de base peut être « perfectionnée » à l'aide de 3 types de « cartouches » :

- Des dictionnaires linguistiques :
  (50 mais les recherches linguistiques sont limitées à 6 langues)
- Des vocabulaires métiers
- Des taxinomies : ce produit propose de « plaquer » 1 à 2 taxinomies sur les réponses d'une requête. Celles-ci peuvent être conçues par le client ou fournies « clés en main » par l'Editeur, qui propose 20 modèles de taxinomies verticales.

La création d'un module destiné à « cartographier » les résultats était en projet en Novembre 2003.

- **ARISEM [26]**

Société française, elle est un des principaux acteurs du marché national. Son chiffre d'affaire était de 3,4 M€ en 2002, racheté récemment par la société Thalès, spécialisée dans la conception de systèmes d'information, au service de l'aéronautique et de l'Armée. Ses clients principaux sont EADS, EDF, France Télécom, le CNES, Pernod-Ricard, etc. « Les technologies Arisem apportent sans doute aujourd'hui des réponses prometteuses en terme d'analyse du contenu et de pertinence de l'information en comparant non pas des mots mais des « concepts universels ». Son nouveau produit **Kaliwatch Server** peut être renforcé par le module **K-Mining,** qui vient « doper » ses fonctions d'analyse et de découverte.

Le module K-Mining permet :

- D'analyser simultanément plus de 1000 documents, quels que soient leurs sources et leurs formats.

- de faire émerger des termes nouveaux non encore répertoriés dans la base de connaissances.

- d'effectuer une surveillance stratégique, de suivre les tendances, de découvrir des données ou de mettre en lumière des relations.

Ce produit intègre une catégorisation automatique censée prendre en compte « le contexte métier » de l'utilisateur (personnalisation), effectuer des recherches sur des corpus « pré-filtrés », et générer des résultats multilingues.

Face au succès de ces nouvelles technologies, de nombreux éditeurs ont décidé de faire évoluer leurs outils traditionnels en y intégrant des processus de traitements sémantiques. D'autres ont décidé de lancer des produits très spécialisés qui peuvent venir compléter des applications plus généralistes (Lingway). Parmi les challengers, on note l'apparition de Temis (spécialiste du textmining), qui vient de s'allier à

Mondeca (spécialiste du Web sémantique) pour mettre au point une plate-forme capable de créer une base de connaissances à partir d'informations brutes issues du web. « *Insight Discover Extractor analyse les sources d'information pour en extraire les concepts métier. Ses règles d'extraction s'appuient sur des règles linguistiques avancées, et sur les ontologies gérées par Intelligent Topic Manager de Mondeca.* » [54-BORDAGE]

BILAN

L'ensemble des points abordés dans cette étude, nous permet d'observer une véritable mutation au sein des processus documentaires, mais aussi en ce qui concerne le rôle des documentalistes.

Si à l'époque du « tout papier », les documentalistes étaient principalement des spécialistes du « contenu » des documents, qui traduisaient les requêtes des utilisateurs et effectuaient les recherches à leur place en s'appuyant essentiellement sur leur mémoire et leur culture personnelle. Aujourd'hui, ils sont devenus des « spécialistes » du traitement de l'information, des concepteurs d'outils permettant de personnaliser l'accès, la diffusion, l'extraction et la présentation des documents pertinents.

Autrefois réservés aux professionnels de l'information, les thésaurus et autres langages d'indexation et d'interrogation se sont substitués aux documentalistes dans leur rôle de médiateur. Pour remplir cette mission de manière performante, les documentalistes ont du transformer et adapter ces outils à ce nouveau contexte d'utilisation. Ainsi, d'intermédiaire unique entre l'utilisateur et les documents pertinents, le documentaliste est devenu un expert en systèmes d'information, dont la mission consiste à collaborer avec les informaticiens et les linguistes en vue de transmettre aux outils de gestion électronique des stratégies de recherche proches de l'esprit humain (par exemple l'exploration par associations d'idées) afin de s'adapter à la variété des profils utilisateurs.

Face à un contexte économique où l'information est devenu un bien de plus en plus stratégique pour les entreprises, les documentalistes ont su adapter et intégrer leurs outils documentaires traditionnels aux processus de gestion électronique de l'information, de manière à satisfaire les exigences de rapidité, d'autonomie et de rapidité des utilisateurs.

Dans la troisième partie de cette étude, nous allons observer comment une grande entreprise a su tirer parti de ces nouvelles technologies pour optimiser l'accès à son patrimoine informationnel.

Partie 3 :

Le portail « Intralignes » d'Air France : un exemple de « corporate taxonomy »

PREAMBULE

Le stage que j'ai effectué cet été, à Air France ne pouvant s'inscrire dans le contexte d'une véritable conduite de projet[48] (pour des raisons d'habilitations administrateur), je n'ai pu participer personnellement à la mise en place d'un portail ou à l'élaboration d'une taxinomie. En revanche, j'ai pu étudier et manipuler au quotidien les taxinomies mises en place sur le Portail du réseau. J'ai été soutenue dans cette démarche par AB, la taxinomiste qui a conçu ces classifications. Par conséquent, mon ambition, ici, est plutôt de vous présenter les raisons techniques, humaines et historiques qui m'ont conduit à suggérer la création d'un portail intranet pour mutualiser les sources d'informations du service DP.SQ. En observant la structure et l'impact des taxinomies déjà en place sur le portail Intralignes, nous étudierons en quoi elles pourraient s'avérer utiles pour le service DP.SQ.

[48] S'il se concrétise, ce projet ne prendra pas corps avant 1 ou 2 ans.

3.1 Air France, une compagnie en quête de transversalité

3.1.1 Contexte d'implantation de l'Intranet[49]

3.1.1.1 le Groupe Air France

AIR France vient tout juste de fêter ses 70 ans, pourtant ses origines remontent au début du siècle dernier et son histoire se fond intimement avec celle de l'aviation internationale. Une histoire longue et complexe, qui éclaire bien des aspects de la culture « Air France » aujourd'hui.

Le groupe Air France est le résultat d'une longue série de fusions, de rapprochements et d'accords de partenariat, tout d'abord entre compagnies françaises, puis avec des compagnies étrangères.

Aujourd'hui le groupe Air France c'est 700 métiers exercés par 70 000 employés répartis dans 85 pays, la fusion avec KLM[50] portant le chiffre des effectifs à 106 500 employés, et l'alliance SkyTeam[51] à 156 000 personnes.

- **Le fusion Air Inter / Air France**

Présente dans l'esprit de tous au sein du groupe, c'est sans nul doute celle qui a encore aujourd'hui le plus de répercussions au sein de l'organisation de la compagnie. La disparition d'Air Inter, englobée par le géant Air France, y est pour beaucoup. Perdre leur identité, a été visiblement difficile à vivre pour le personnel de cette compagnie aux dimensions plus humaines. Par ailleurs, des différences de statut, entre le personnel Air France et les ex-air Inter contribuent à maintenir le clivage et rend difficile l'harmonisation des méthodes de travail et la circulation de l'information au sein de l'entreprise.

[49] Je reprends ici des passages de mon rapport de stage rédigé en février, à l'issu des premières semaines passées au sein de l'entreprise.
[50] Compagnie nationale Néerlandaise
[51] Accord de partenariat avec des compagnies étrangères

Depuis 2000, la naissance de l'alliance SkyTeam, via la signature d'un accord de partenariat entre Air France , Delta Airlines, Aéromexico et Koréan Airlines, scelle l'émergence d'une nouvelle ère, qui atteint aujourd'hui sont apogée avec la fusion Air France/KLM

Extrait du discours du 5 mai 2004 prononcé par les PDG d'Air France (Jean- Cyril Spinetta) et de KLM (Léo Van Wijk) :

« *Air France et KLM ont décidé de joindre leurs forces et leurs destins pour construire une entité nouvelle à la dimension du grand marché européen. Aujourd'hui, nous partageons une ambition : faire partie des quelques compagnies assez puissantes pour jouer un rôle de leader au sein des alliances globales qui structureront à l'avenir le transport aérien. Autrement dit, être en mesure de jouer les premiers rôles sur la scène mondiale, afin d'améliorer notre rentabilité, offrir le meilleur service à nos clients et pérenniser nos emplois. Les hommes et les femmes d'Air France et de KLM qui vont construire le nouveau groupe peuvent s'engager dans cette voie avec fierté* ».

Le 22 février 2005, c'est la consécration, le groupe déjà n°1 mondial du transport aérien, est consacré meilleure compagnie aérienne de l'année par le journal américain « Air transport World », qui fait autorité dans le milieu de l'aviation pour sa rigueur et son professionnalisme.

3.1.1.2 La Compagnie Air France

Carte d'identité de la compagnie

1ère compagnie Européenne pour le nombre de passagers transportés

3ème rang mondial pour le transport international de passagers

4ème rang mondial pour le transport international de fret

2ème prestataire mondial d'entretien aéronautique

1ère Major européenne en terme de part de trafic

L'activité de la compagnie est centrée autour de 4 grands pôles :

- Le transport de passagers
- Le transport de fret (marchandise)
- La maintenance des avions (prestataire de service pour d'autres compagnies)
- La formation

La gestion de ces activités est assurée par onze départements, situés sous l'autorité de la Présidence et du Siège qui sont implantés à Roissy :

- Air France Cargo
- Commercial France
- Commercial International
- Coordination
- Economie et Finances
- Exploitation Sol
- Industriel
- Marketing et Réseau
- Opérations Aériennes
- Ressources Humaines
- Systèmes d'Information

Des départements répartis entre les différents « sites » de la compagnie :

- Roissy
- Paray-vieille-Poste
- Le Bourget
- Toulouse
- Orly
- Vilgénis
- Valbonne

3.1.1.2.1 Un éclatement géographique, source de difficultés techniques et relationnelles

Si en apparence la structure de ces pôles semble très « logique », il n'en n'est pas de même de leur répartition géographique. Ainsi, le pôle exploitation sol, chargé de la maintenance des avions, voit certains membres de son personnel obligés de « migrer » » une, voir plusieurs fois par semaine vers Roissy pour des réunions se tenant au siège. Ou pire encore, pour récupérer des documents nécessitant des

mises à jour hebdomadaire, mais dont le poids rend l'envoi via le réseau impossible[52].

Par ailleurs, au sein d'un même pôle, les services d'exploitation peuvent être localisés sur deux ou trois sites alors que sa direction sera implantée dans un quatrième (le plus souvent à Roissy), ce qui complique les rapports avec la hiérarchie.

On perçoit ici, l'importance d'un système informatique cohérent, uniformément réparti et fonctionnel. Sans cette structure technique, l'information loin de pouvoir être diffusée à tous, rapidement et de manière fiable, trouve son chemin en empruntant des voies parallèles, où les rumeurs, les blocages et les malentendus prennent naissance. Cette « politique », consciente ou non, de rétention et de morcellement de l'information, contribue à renforcer les corporatismes et freine l'intégration des nouveaux arrivants (salariés ou partenaires commerciaux) dans la compagnie. Pourtant la compagnie affiche aujourd'hui, un véritable engagement pour accentuer la transversalité dans l'entreprise, mais les veilles habitudes ont la vie dure et, loin du siège, le message semble plus difficile à faire passer.

3.1.1.2.2 Une organisation interne qui manque encore de transparence

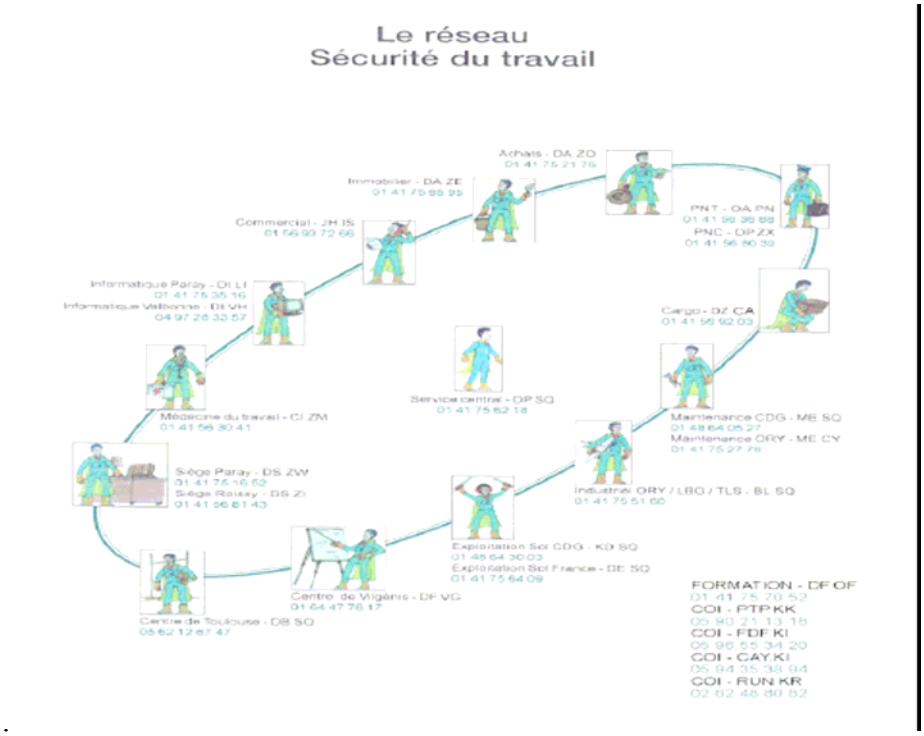
De la désignation des services par des sigles, à l'organigramme (qui ne propose que les niveaux les plus hauts de la hiérarchie), tout contribue à maintenir une sorte de flou sur l'organisation interne de la compagnie.

Par ailleurs, il n'existe aucun organigramme détaillé par pôles, et ce malgré les demandes répétées du personnel, qui semble éprouver de grandes difficultés à visualiser les responsabilités et la place de chacun au sein du groupe.

Cette situation n'est pas sans conséquences pour le service DP.SQ, dont le réseau est par essence transversal, puisqu'il regroupe des employés appartenant à tous les

[52] On constate des variations géographiques allant de 64kb à 5Mb. Certaines transactions (en fonction des postes) étant limitée à 20kb.

services de la compagnie et dispersés sur l'ensemble du territoire (France et outre-mer).



.

Figure 6 : organigramme fonctionnel du réseau sécurité du travail

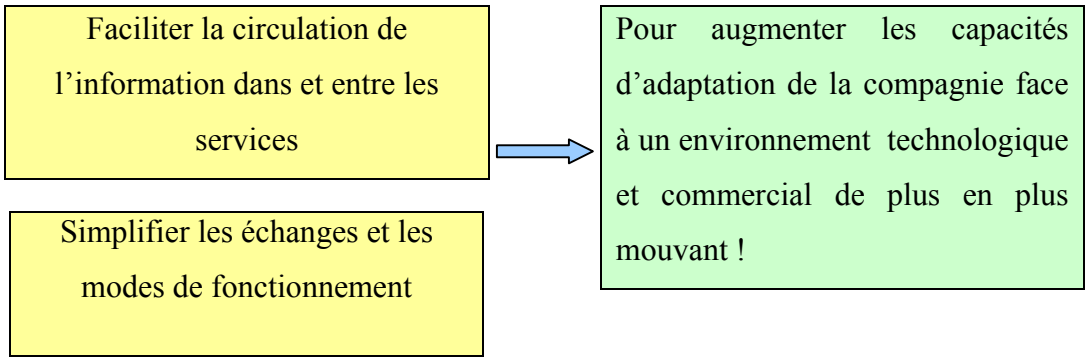
Autant de cultures d'entreprises, de méthodes de travail, d'infrastructures informatiques, techniques et administratives différentes, ne sont pas sans générer des problèmes de fonctionnement. D'où l'importance accordée à la mise en place d'un média permettant à l'ensemble des salariés du groupe, d'accéder aux informations qui lui sont nécessaires pour le bon exercice de ses missions, quelque soit sa localisation géographique. Le déploiement du réseau Intranet et de ses portails « métiers » a été officiellement lancé en 2002, il s'inscrit dans la ligne directrice des projets d'entreprise Major[53].

[53] Désigne les projets qui s'inscrivent dans la nouvelle dynamique d'équilibre d'AF : gagner la préférence des clients AF, jouer de leur Atouts, agir ensemble, être rentable pour financer l'avenir.

3.1.2 Quel intranet pour Air France ?

En 2002, au moment du lancement du projet, Air France comptait environ 120 sites spécialisés et 6000 bases Lotus Note, et les demandes de création continuaient à affluer. Décidée à réguler le phénomène, en optimisant le partage et l'accès aux contenus déjà publiés, la direction a lancé son programme INTRANET (voir annexe 3) pour structurer et mutualiser les informations produites par la compagnie.

Le but de ce programme est avant tout de fournir une structure simple, une méthode commune, et des outils pratiques aux services désireux de se doter d'un portail. Cette démarche vise aussi un objectif plus large qui est l'amélioration de la transversalité au sein de la compagnie :



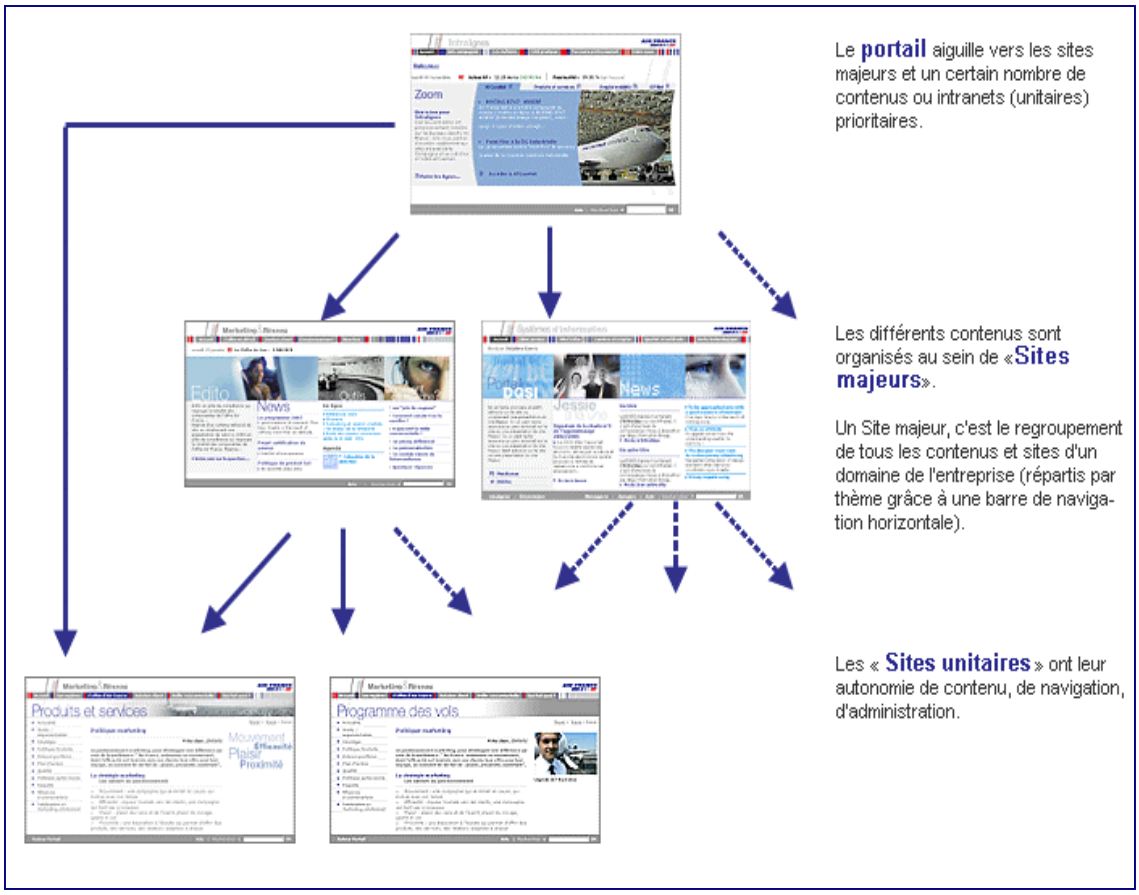
3.1.2.1 Démarche et structure

Les informations rapportées dans cette partie et la suivante (3.1.2.2) s'appuient et s'inspirent de l'étude menée par Melle Audrey Blanchard, Taxinomiste à Air France, lors de son stage de DESS au sein du service Intranet de la compagnie. [88 – BLANCHARD]

- Proposer une structure commune pour tous les sites afin de garantir la cohérence du portail
- Structurer l'existant (demander au services de procéder au « nettoyage » des « containers[54] » qui seront indexés par le moteur Verity et consultables via le portail)
- Limiter le texte, privilégier la mise en ligne des processus, intégrer les applicatifs avec un souci de fiabilité, de gain de temps, de proximité.
- Limiter l'Intranet à l'indispensable et à l'utile
- Créer la structure d'accueil des sites actuels et des futurs sites
- Proposer un socle technique unique
- Permettre aux décideurs de piloter selon les dimensions coût, qualité, délais...
- Proposer un accès sécurisé et personnalisable (en fonction du corps de métier des utilisateurs)

La démarche adoptée, consiste à mettre en place un portail général, à partir duquel les utilisateurs pourront accéder à des portails métiers : les sites majeurs, qui donneront eux-mêmes accès à des sites spécialisés : les sites unitaires.

[54] Par containers, on entend tous les structures qui contiennent des documents électroniques.



Dans un premier temps, la compagnie a sélectionné les 7 départements, les plus susceptibles d'offrir des informations et des services qui permettront un vrai gain de temps aux salariés.

- informatique,
- communication,
- formation,
- Personnels Navigants,
- commercial,
- marketing et réseau,
- achats

Ces sites sont structurés de manière à fournir un accès personnalisé aux informations en fonction du public qui les consulte. Il est donc conçu pour être d'une part un outil « grand public »[55], et d'autre part un outil d'aide à la décision pour les cadres dirigeants de l'entreprise.

Figure 8 : interface du site majeur du département ACHAT

Petit à petit des sites unitaires, correspondant à des pôles d'activités très spécifiques sont intégrés dans les sites majeurs des directions dont ils dépendent. Ainsi le portail du service Sécurité et Conditions de travail, trouverait normalement sa place au sein du site majeur des Ressources Humaines (à l'instar du service médical présenté dans l'exemple suivant, fig. 9), mais le futur déménagement du service à Roissy, incitera peut-être la direction à regrouper la sécurité du travail, la santé et la qualité dans un même pôle.

[55] Dans le contexte de la compagnie AF, grand public désigne l'ensemble des salariés de la compagnie à l'exception des instances dirigeantes.

Figure 9 : exemple de site unitaire au sein du site majeur des ressources humaines

Dans un premier temps, tous les sites disposeront de la même interface, puis au fur et à mesure de l'extension du réseau et de l'intégration des autres sites majeurs (et de la Wébisation des bases lotus qui pourront elle aussi être insérées dans les portails), les sites deviendront personnalisables (dans le respect de la charte graphique d'Air France). Le but ultime étant de proposer à chaque utilisateur un bureau virtuel, où ils disposeraient de toutes les sources d'information et de tous les outils qui lui sont utiles au quotidien. De cette manière, on pourra mutualiser les applications validées par l'entreprise, ce qui limitera les coûts et harmonisera les outils de travail. De plus, en « dissimulant » toutes les applications[56] derrière une seule page d'accès évolutive, la compagnie limite les changements incessant d'interfaces, ce qui permet d'optimiser, au quotidien, le travail des salariés.

[56] Voir Fig.8, dans la rubrique outils, on remarque un accès à la base Iso Target, qui intègre un processus de workflow pour la publication des documents internes « stratégiques »

Figure 10 : exemple de bureau personnalisé via Intralignes

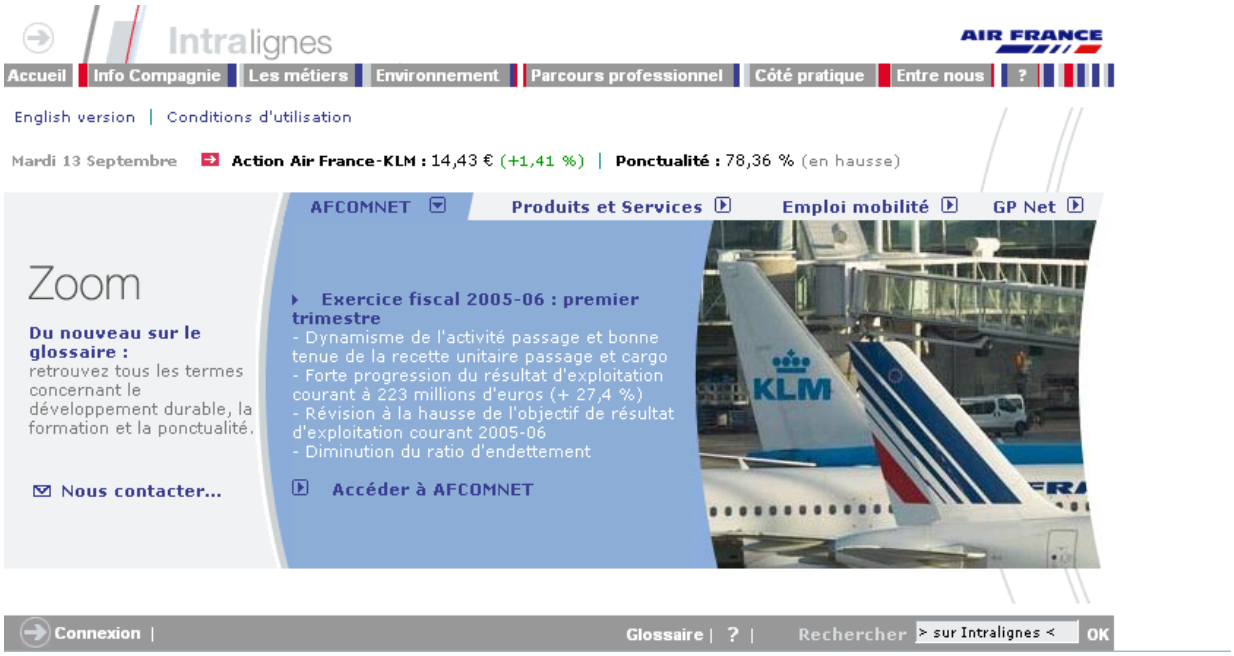
3.1.2.2 Le portail Intralignes

3.1.2.2.1 Structure

Le squelette du portail repose sur 5 grandes thématiques :

- ✓ **Info Compagnie**
  L'actualité et les informations institutionnelles de la Compagnie.
- ✓ **Les métiers**
  Les sites des grands métiers et Directions de la Compagnie.
- ✓ **Côté pratique**
  Des services et informations qui ont pour objectif de faciliter la vie au quotidien dans l'entreprise
- ✓ **Parcours professionnel**
  Les informations concernant la gestion de votre carrière, les opportunités et l'évaluation des compétences.
- ✓ **Entre nous**
  Les services proposés aux personnels d'Air France.

Figure 11 : page d'accueil du portail Intralignes - septembre 2005



3.1.2.2.2 Navigation

La navigation sur le portail se fait sur deux axes :

- Horizontalement, au travers des 6 rubriques précitées. Ces dernières donnant accès à des menus déroulants qui permettent d'accéder à des sites majeurs ou unitaires. Dans chacun des sites associés, la barre de navigation horizontale reprend cette division en 6 rubriques, qui cette fois, correspondent à des thématiques spécifiques au métier exploré.

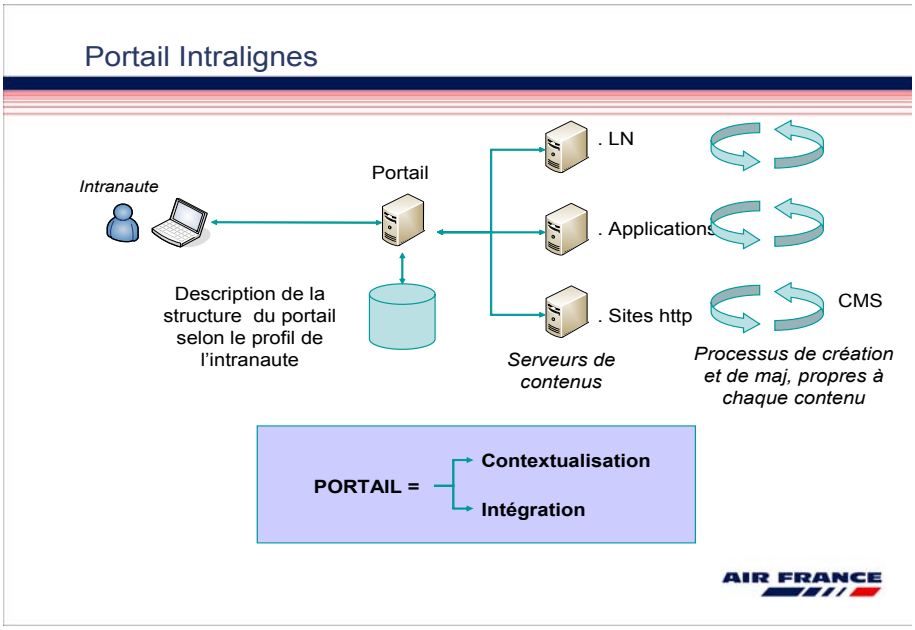
- Une navigation intuitive, dite de « découverte », qui s'effectue par l'intermédiaire de 6 onglets, proposés au centre de l'écran. La démarche adoptée, consiste à mettre en valeur des rubriques « pratiques », régulièrement consultées par les salariés. En les plaçant au centre du portail,

on raccourcit le cheminement vers le site majeur qui les abrite. Faciles à localiser, immédiatement exploitables par le salarié, elles contribuent à fidéliser les « intranautes » d'AF, en vue d'introduire chez eux un réflexe « Intralignes », au même titre qu'il existe un réflexe « Google » chez les Internautes.

3.1.2.2.3 Gestion des contenus et processus de recherche

Le portail Intralignes a été développé dans le cadre de l'implantation du logiciel Verity. N°1 du marché, ce logiciel est plébiscité pour sa capacité à explorer et à indexer des structures composées de sources d'informations hétérogènes, tant du point de vue de la forme (format de fichier, technologie des infrastructures de stockage), que du point de vue du fond (champs sémantiques multiples, multilinguisme, documents de vulgarisation et documentation technique, etc..).

Figure 12 : schéma structure technique du portail intralignes d'AF



On accède au moteur de recherche par un lien (situé dans le footer[57], il devrait bientôt être déplacé, cet emplacement étant « ergonomiquement » catastrophique), qui donne accès à l'interface de recherche. Parallèlement, des « portlets[58] » de recherche sont introduits progressivement dans les pages d'accueil des sites majeurs et unitaires.

Le portail Intralignes d'air France propose 3 types de démarches :

- Une recherche simple
- Une recherche avancée
- Une recherche thématique par navigation dans une arborescence (la Taxinomie de l'entreprise)

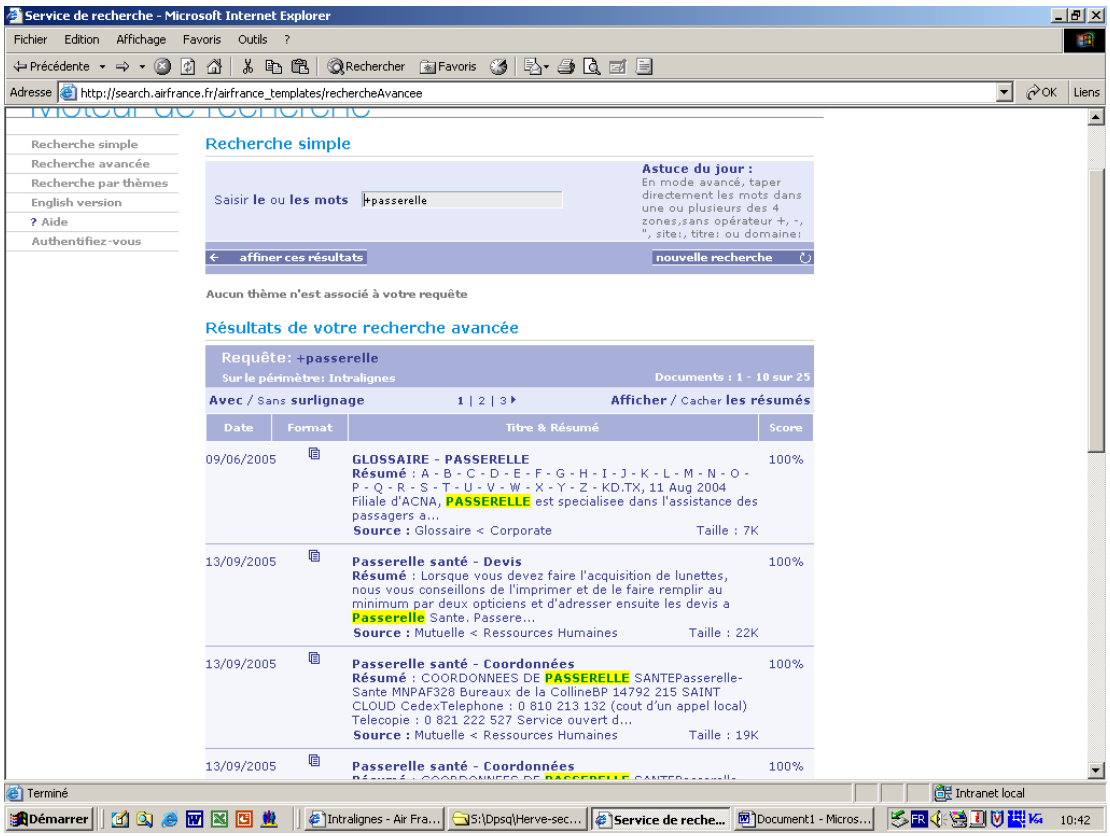


Figure 13 : Interface recherche simple sur Intralignes AF

[57] Terme qui désigne une barre de taches, située en bas de la page web
[58] Mini interface de recherche insérée dans la page d'un site, en général la page d'accueil.

On remarque que le terme recherché, est ensuite signalé dans les résultats par un surlignage jaune, afin que le lecteur puisse le repérer plus rapidement et vérifier s'il est employé, dans le contexte sémantique souhaité.

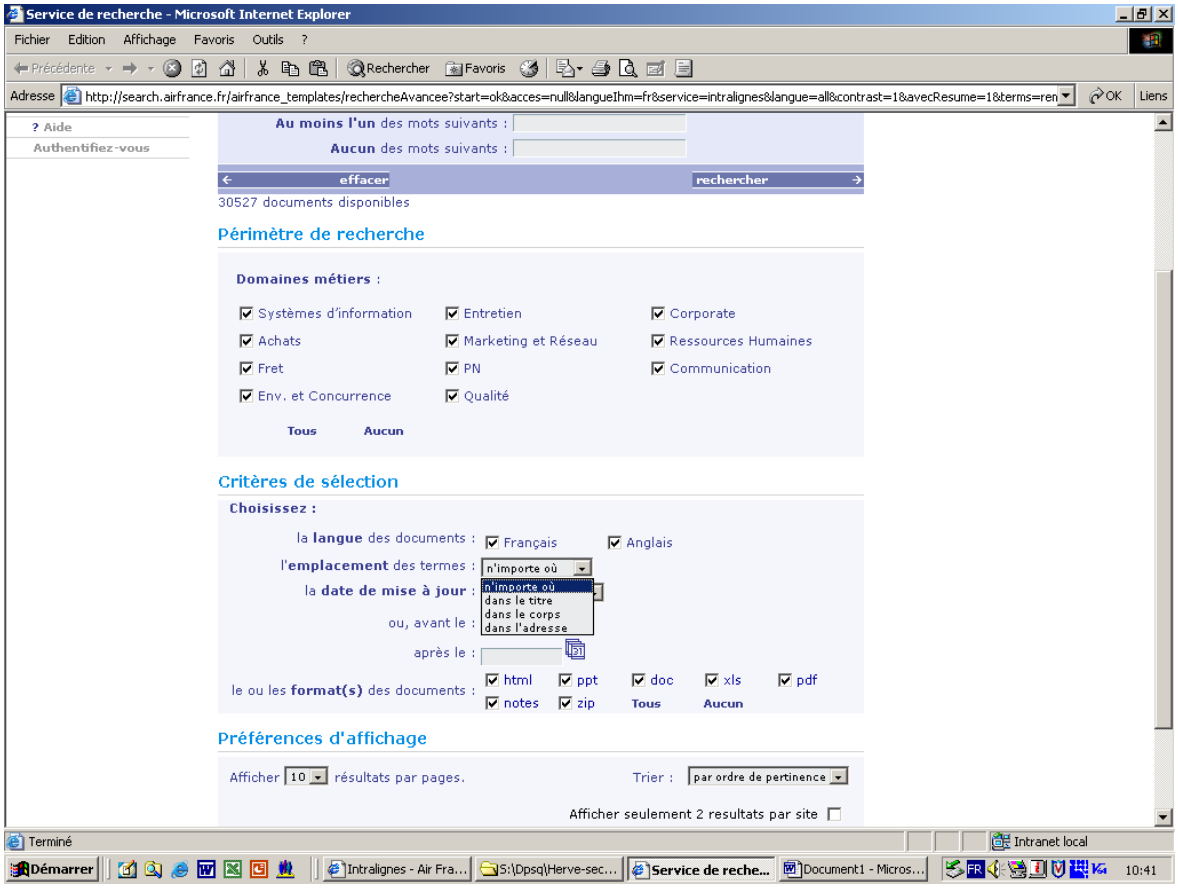


Figure 14 : interface recherche avancée portail Intralignes AF

La recherche avancée propose de limiter le périmètre de recherche en cochant sélectionnant des domaines métiers, qui correspondent aux grandes directions de la compagnie telles quelles existent dans l'organigramme hiérarchiques. Les fonctionnalités sémantiques et linguistiques du moteur Verity, sont perceptibles sur cette interface, puisqu'il existe un critère de sélection par l'emplacement des termes (pondération) et par la langue.

Figure 15 Interface de recherche par thème Intralignes AF

On remarque que la thématique proposée ne colle plus à l'organigramme hiérarchique. Il s'agit ici de proposer une représentation fonctionnelle et pragmatique de l'entreprise, plus proche des besoins des utilisateurs.

Chaque thématique est déclinée en sous-thèmes qui permettent d'accéder directement aux documents indexés et classés automatiquement dans cette catégorie par le moteur Verity.

Le Portail Intralignes se veut avant tout un outil pratique, destiné à faciliter l'accès à des informations professionnelles spécifiques à chaque corps de métiers. Il intègre un système d'accès sécurisé, qui repose sur le module de gestion des profils utilisateurs proposé par le logiciel Verity. Ce système s'appuie sur le nouveau processus d'immatriculation des salariés dans l'annuaire d'entreprise (système SAP-CORHUM), et permet d'identifier et de rassembler sous un seul login (et un seul mot de passe) toutes les accréditations d'un salarié. C'est sur ce même principe que fonctionne le module Verity Intelligent Classifier, qu'Air France est en train d'implémenter. Cette option permettra de personnaliser la présentation des résultats d'une recherche, en fonction des centres d'intérêt et du vocabulaire professionnel de utilisateur.

3.2 Les taxinomies : des outils au service de la pertinence

3.2.1 Les Taxinomies : un « nouveau » mode de recherche pour les portails Intranet

3.2.1.1 Conception d'un portail, une création en mode projet

L'élaboration et la mise en place des taxonomies font partie intégrante de la conception d'un portail conçu dans une infrastructure gérée par le logiciel Verity K2.

La création d'un portail est une opération menée en mode projet, par conséquent elle nécessite la constitution d'une équipe, l'identification des divers partenaires et la mise en place d'un planning pour coordonner les étapes de l'opération.

Ce type de projet, implique une étroite collaboration entre le maître d'oeuvre du portail (la direction métier ou le service qui a passé la commande), les concepteurs du produit (le plateau Intranet), les développeurs (service informatique situé à Toulouse), et l'Assistant maîtrise d'ouvrage (AMO) qui assure l'interface entre ces différents acteurs.

Le portail Intralignes d'Air France et son système de recherche Verity, n'échappe pas à la règle. Leurs mises en place a été gérées par l'équipe Intranet, qui se trouve au cœur de la conduite de projet (cf. annexe 2) pour veiller à la bonne coopération des autres intervenants. Ce service est là pour s'assurer de la cohérence des différents portails, par rapport au programme établi pour le réseau Intranet (respect de la charte graphique, des principes de navigation, etc..).

Mon propos n'étant pas de décrire en détail toutes les étapes de la mise en place du portail Intralignes d'Air France, je vous propose de les découvrir au travers d'une représentation graphique du planning suivi par l'équipe.

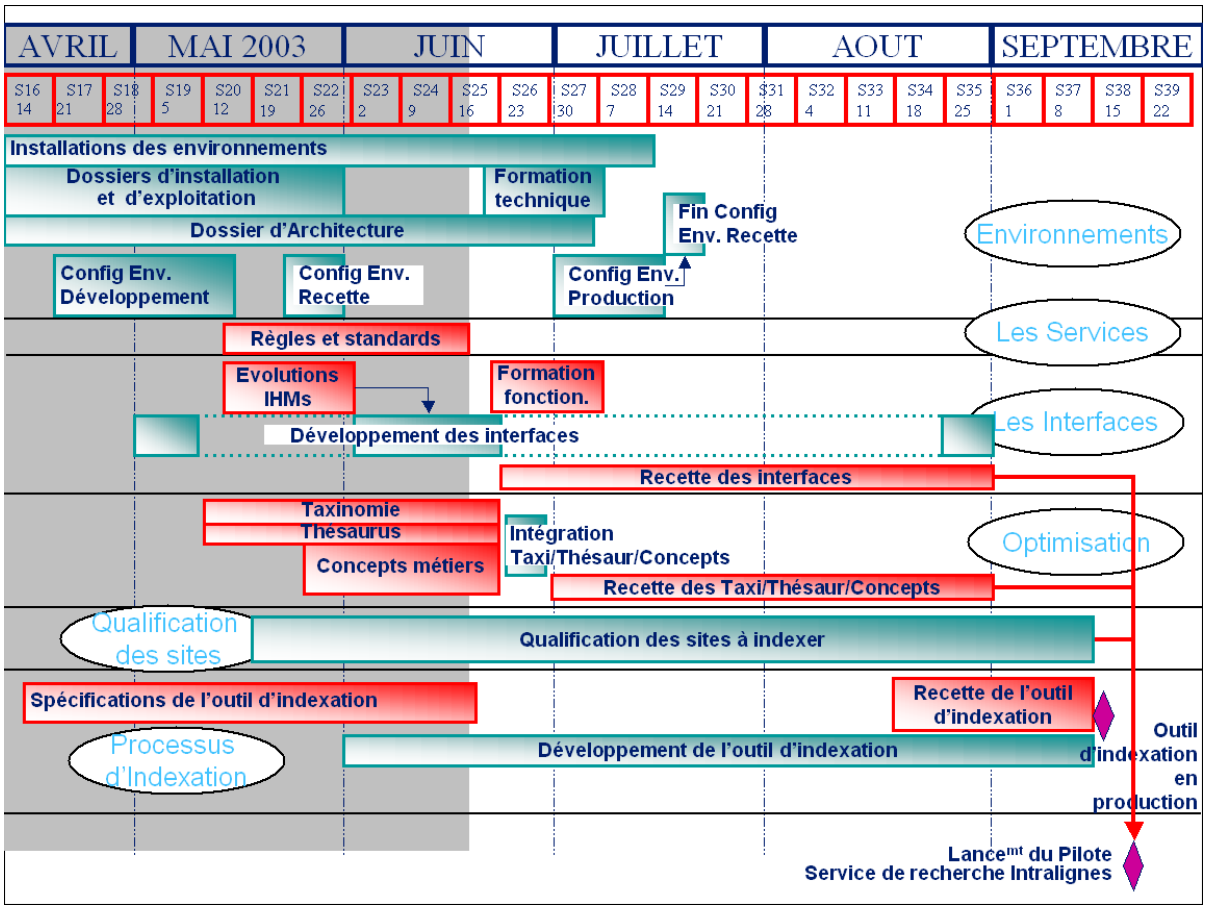


Figure 16 : Planning création du système de recherche du portail Intraligne – Audrey Blanchard, rapport de stage InnoVinfo, 2003

3.2.1.2 Rédaction et intégration des taxinomies

Outre les performances du moteur, ce sont surtout les opportunités offertes par les taxinomies qui ont poussé la compagnie Air France à porter son choix sur le logiciel Verity. Si elles présentent de grands atouts pour les entreprises, elles demandent aussi beaucoup de rigueur.

3.2.1.2.1 Pourquoi choisir d'utiliser des taxinomies à Air France ?

- Pour améliorer la pertinence des résultats
- Pour fournir des outils de recherche plus intuitif
- Pour éliminer les ambiguïtés de vocabulaire
- Pour contextualiser les informations

3.2.1.2.2 Comment les taxinomies permettent-elles d'obtenir ces résultats ?

[87-BLANCHARD]

- En dressant une cartographie des connaissances et des concepts liés à la compagnie AF et aux différents métiers qui y sont exercés.
- En proposant un système de guidage aux utilisateurs du portail, via une navigation hypertexte, qui permet d'évoluer dans l'arborescence de la taxinomie, en affinant petit à petit la sélection des thématiques de recherche. (processus de l'entonnoir)
- En associant les fonctionnalités d'un thésaurus aux taxinomies, pour offrir aux utilisateurs un système de suggestion qui affiche les concepts voisins de ceux de la requête (le « Voir aussi » des thésaurus)
- En associant la taxinomie au système de recherche par champ d'interrogation, de manière à restreindre le périmètre de recherche au secteur qui intéresse l'utilisateur.

3.2.1.3 Des conditions nécessaires au maintient de la pertinence

- Mettre régulièrement à jour les thématiques et les concepts métiers.
- Associer les différents métiers à la maintenance et à l'évolution de la cartographie de la compagnie.
- S'assurer de la cohérence et de la logique du classement des documents dans la taxinomie, en limitant le nombre d'intervenant chargé de cette tâche
- Sensibiliser les « contributeurs » (les producteurs d'information, qui alimenteront le site en contenu) au processus de gestion des connaissances (KM). En insistant sur la nécessité de « décrire » les documents produits, de manière à optimiser le référencement du fond.

3.2.2 Création et mise en place d'une taxonomie fonctionnelle de la compagnie Air France

3.2.2.1 Conception, rédaction et intégration des taxinomies dans le logiciel de recherche

Dans le planning précédent (fig.16) nous pouvons voir que leur phase de rédaction s'est étendue sur **deux mois**, et que leur introduction dans le système de recherche du portail, correspond à la phase d'optimisation des processus. C'est-à-dire peu de temps avant la mise en ligne du pilote.

Pour dresser cette « cartographie des connaissances » de la compagnie, AB, est allée à la rencontre de toutes les directions « métiers » d'Air France (à ce jour 13). Elle a établi avec eux une carte fonctionnelle de leur activité.
Dans un second temps, ces taxinomies ont été organisées et harmonisées en collaboration avec la direction des ressources humaines, pour aboutir finalement à une classification générale composée de 26 thématiques « racines ». Chacune de ses

« racines » est à l'origine d'une arborescence qui reflète les domaines d'intervention et l'identité du métier exploré.

Figure 17 : exemple d'arborescence pour la thématique Air France Corporate

Une fois validée par chaque entité, les taxinomies sont introduites dans le logiciel de gestion de contenu.

3.2.2.2 Des dispositifs pour optimiser la construction et l'exploitation de ces classifications :

- Les business Rules (ou Classify Rules), se sont des règles d'inférences qui permettent d'attribuer à chaque rubrique « racine » un certain nombre de propriétés.
- Les topic set, se sont des concepts qui sont combinés pour « décrire » les rubriques. Ces combinaisons permettent d'établir des requêtes types qui augmente la pertinence des recherches effectuées à partir des rubriques.

- La création des catégories, peut aussi être réalisée automatiquement à l'aide d'une technologie propriétaire LRC (Logistic Regression Classifier). Dans ce cas, le logiciel s'appuie sur des échantillons de documents étiquetés pertinents ou non pertinents.

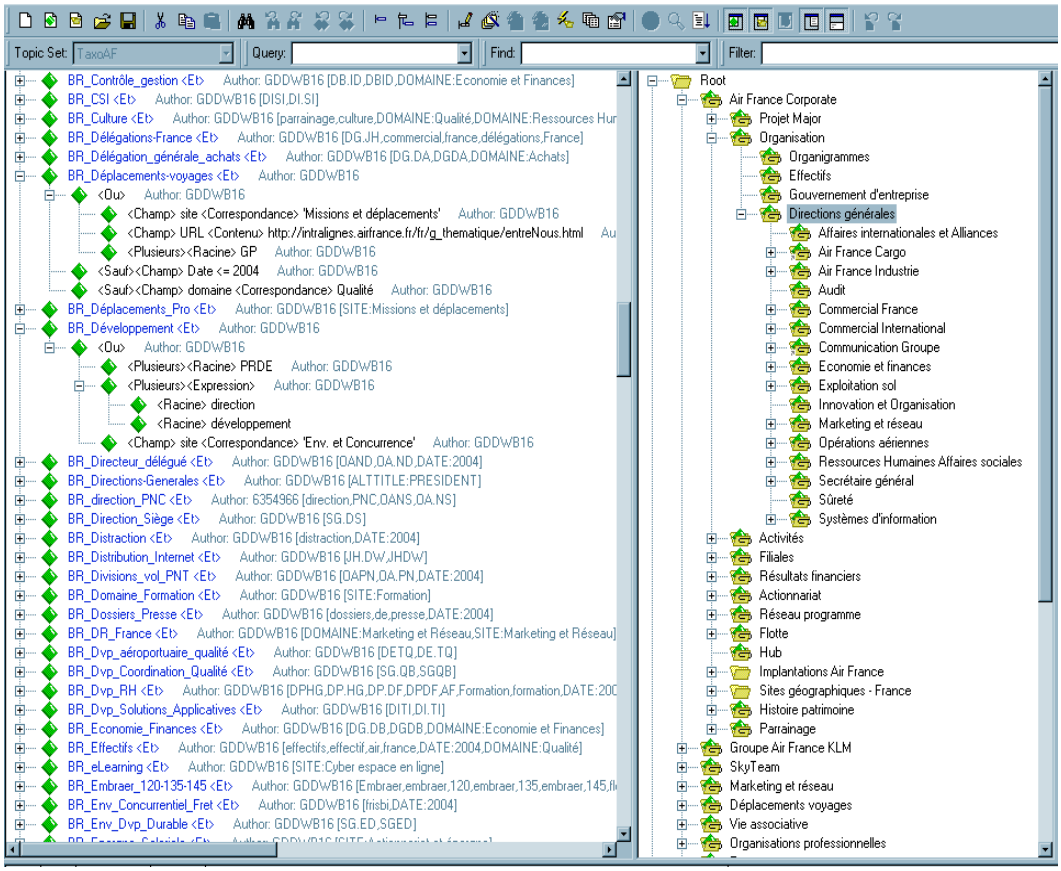


Figure 18 : Interface de gestion des taxinomies dans Verity K2

La mise en place des taxinomies s'inscrit dans une politique de gestion des connaissances, par conséquent leur usage est optimisé si les documents sont décrits et indexés à l'aide de la taxinomie avant d'être publié sur le réseau. C'est pourquoi, parallèlement à la construction du portail, AF a mis un outil de publication à la disposition des salariés. Ce dernier permet d'inclure dans les documents, des champs de description obligatoires (métadonnées) qui peuvent être alimentés à l'aide de la taxinomie ou de listes de mots-clés. Cette démarche contribue à améliorer le référencement du fond.

3.2.2.3 L'impact des taxinomies sur la compagnie
[87-BLANCHARD]

Bénéfices fonctionnels

- Un outil de recherche intuitif, spécialement conçu pour les « intranautes »
- une meilleure perception des activités de la compagnie
- une circulation de l'information plus facile grâce au décloisonnement des métiers
- Un accès plus homogène aux projets corporate

Bénéfices techniques

- Une augmentation et une amélioration de l'indexation des documents publiés sur le réseau
- Homogénéisation des formulaires de saisie dans l'outil CMS

Bilan d'exploitation

Après 1 an d'exercice, 20 % des recherches effectuées par les intranautes sur le portail Intralignes sont réalisées à l'aide des thématiques.

L'implantation récente de nouveaux sites unitaires va sans doute drainer de nouveaux utilisateurs, l'enrichissement permanent de la taxonomie contribue sans aucun doute à l'intérêt que leur porte les salariés, qui disposent ainsi d'un outil de plus en plus fin.

3.2.3 Un site intranet pour le Service Sécurité et Conditions de travail

Le service sécurité et conditions de travail est constitué d'un service central (D.P.SQ) et d'un réseau de 500 représentants (relais, conseillers, CHSCT, médecins, inspecteurs du travail) avec lesquels il entretient des relations fonctionnelles.
Le réseau s'étend sur tous les sites d'Air France, y compris dans les Dom-Tom, partout où il y a des employés d'Air France, il doit y avoir des relais sécurité pour transmettre les recommandations du service central.

Cette **hétérogénéité** ne s'arrête pas au domaine géographique. Issus de tous les corps de métiers, certains représentants du réseau exercent leurs fonctions sécurité à plein temps ou à temps partiel (parfois 2H de permanence / semaine) en plus de leur profession. Leur appartenance à des professions techniques, exercées sur le terrain ne leur donne pas toujours accès aux mêmes infrastructures bureautiques. Contrairement au service DP.SQ, leur cœur de métier n'est pas la recherche d'information. Leurs besoins sont souvent ponctuels, mais en revanche ils sont très souvent liés à une notion d'urgence (accident, nouveau matériel, etc…).

Le fait que les membres du réseau et du service central, dépendent hiérarchiquement de plusieurs directions, rend très difficile l'harmonisation des outils et des méthodes de travail. Cette situation se traduit pour l'ensemble de la compagnie par :

- Une disparité du matériel informatique (y compris au sein d'un même service)
- Une mauvaise signalisation des bases existantes,
- Un accès aux sources d'information électroniques limité,
- Un manque d'homogénéité dans la répartition des accès à Internet,
- Des difficultés pour acquérir de nouvelles applications informatiques (logiciel Access, pdf writer, etc …), car elles doivent nécessairement être validées et référencées par la direction de achats et les services informatiques de AF

- Un budget documentaire élevé. En l'absence d'un service central des achats documentaires, qui permettrait de mieux gérer les abonnements et de limiter leur coût, on constate une redondance inutile de certains abonnements, alors que le budget manque pour acquérir des produits parfois un peu onéreux.

3.2.3.1 Une nécessaire mutualisation des outils et les sources d'information

L'audit des sources d'information réalisé durant cette mission, a permis d'établir une typologie des sources d'informations utilisées et produites au sein du service.

Au sein du service central, on note un fort pourcentage de sources d'informations électroniques, principalement Internet (le cédérom des éditions législatives « circulant » dans le service). Ensuite, viennent les journaux spécialisés et les ouvrages de références en matière de droit du travail. En dernier lieu, on trouve quelques bases électroniques Internes.

Au sein du réseau la situation est quelque peu différente, certains représentants ne disposant pas d'une habilitation pour utiliser Internet, ils effectuent une partie de leur recherche chez eux (leurs sites « favoris » sont le plus souvent les mêmes que ceux des préventeurs du service central). Par conséquent, leurs d'informations sont le plus souvent issues de la presse spécialisée, des bulletins électroniques envoyés par HM, et des ouvrages de références.

La redondance de certaines sources d'information est bien entendu inévitable dans un service aussi spécialisé. Les sites officiels comme l'INIST, ou Légifrance sont incontournables et connus de tout le personnel du service. En revanche, la spécialisation d'un préventeur ou d'un ergonome dans tel ou tel domaine, l'amène à découvrir des sites ou des ouvrages qu'il pourrait être intéressant de conserver et de partager avec d'autres collègues pour une utilisation ultérieure. D'où l'intérêt, d'un site où l'on pourrait proposer aux utilisateurs une bibliothèque des meilleurs sites d'information, régulièrement mis à jour.

3.2.3.2 Mais pourquoi un site sur le portail intralignes, plutôt qu'une base Lotus ?

L'information traitée et produite par le service DP.SQ est essentiellement technique et juridique. Ces deux domaines nécessitant une grande précision des termes, il est absolument indispensable de les **contextualiser** au moment de la requête, afin de lever toutes les ambiguïtés susceptibles de générer des résultats erronés.

Compte tenu de la structure complexe du service et de la variété des cultures professionnelles qui s'y côtoient, il me semble indispensable de mettre en place un système d'indexation des documents, susceptible de **satisfaire toutes les approches, quelque soit le profil de l'utilisateur.**

La contextualisation des termes peut être obtenue grâce à l'intégration d'une taxinomie reflétant l'organisation fonctionnelle du service. Cette classification est ensuite insérée dans le système de recherche du portail (Verity). L'ajout de **glossaires métiers** permet ensuite d'enrichir l'indexation et de « **personnaliser » le classement des résultats**, ce qui a pour effet d'optimiser le succès de la recherche.

Les bases Lotus ont pour principal inconvénient de ne pas pouvoir être consultée à distance, à l'inverse des sites intranet qui offrent une possibilité d'accès (avec login) via l'extranet de la compagnie. Par conséquent le partage des informations avec des représentants éloignés géographiquement ou travaillant en horaire décalé n'est envisageable qu'à la condition de disposer d'un portail sur Intralignes.

Ce site devra remplir quatre missions :

- Etre un outil de partage des connaissances,
- Mutualiser les applications et les logiciels de publications
- Personnaliser les accès à l'information en fonction des cultures professionnelles,
- Soutenir le service central dans sa mission de prévention des risques professionnels, en ouvrant un espace de communication direct avec le « grand public » d'Air France.

Conclusion

« *If not already learning how to use taxonomies, enterprises should start learning now. They are not an experimental technology but key to organizing, accessing and managing corporate assets*." [49-PLOSKER, p. 59]

Leurrées par le mythe d'un web omniscient, beaucoup d'entreprises ont dissous leur centre de documentation, et reporté les processus de recherche sur leurs employés. Mais sans l'intermédiaire des spécialistes de la gestion de l'information et surtout sans les outils qu'ils concevaient pour localiser physiquement les documents en fonction de leur contenu, les réseaux d'entreprise sont devenus des jungles hypertextuelles difficilement gérables. Des espaces non structurés où les salariés prennent le risque de se perdre dans un labyrinthe de pages (lost in hyperspace syndrome) [39- VILLANOVA OLIVER]. Longtemps gérés par des informaticiens, ces intranets ont avant tout axés leur développement sur l'ajout de fonctionnalités sans cesse plus performantes (base de données « webisables », processus de workflow, groupware, visioconférences), mais pas sur l'amélioration de l'accès aux documents.

Or, dans une entreprise de 70000 employés (comme Air France), même en réduisant la production de documents « publiables » à 1 fichier pour 2 cadres par semaine, le nombre de documents disponibles sur un Intranet devient vite astronomique[59]. Avec une telle masse d'information, les résultats obtenus suite à une recherche en texte intégral deviennent absolument ingérables. En effet, même si la liste des réponses contient à coup sur des documents pertinents, ils sont perdus au milieu de dizaines d'autres résultats, qui comportent bien le terme recherché mais dans un contexte d'utilisation différent. Et jusqu'à l'arrivée récente des moteurs de recherche linguistique et sémantique, rien ne permettait de regrouper ces résultats selon des critères de pertinence contextuels. Sans cette technologie, il devient impossible de retrouver tous les documents pertinents, puisqu'ils sont dispersés sur des dizaines de pages de résultats. On imagine la perte de temps que cela représente pour les employés, qu'ils soient spécialistes de l'information ou pas.

[59] On compte plus de 13000 documents rien que pour la thématique « racine » : Air France Corporate. Sachant que toutes les directions et tous les services qui en dépendent, ne disposent pas encore de vitrine sur l'Intralignes, le volume réel des documents rattachables à cette catégorie est sans nul doute largement supérieur.

C'est pourquoi aujourd'hui, les entreprises sont dans l'obligation d'apprendre à (re)construire leur univers informationnel, en commençant par « classer » leurs informations, pour mieux les contextualiser. Pour faire face à cette demande, les logiciels de gestion de contenu ont commencé à se doter de nouveaux outils, aux origines et aux appellations volontairement obscures. Il semblait, en effet, difficile de séduire des clients en perpétuelle quête de modernité par l'emploi de terme vu et revu, et extrêmement connoté « documentation ». C'est pourquoi, on a vu peu à peu surgir dans les plaquettes des éditeurs, des termes comme Ontologies ou Taxinomies.

Après examen, on pourrait être tenter de dire que ces outils aux noms « ronflants », ne sont finalement que des langages classificatoires, intégrant des relations plus ou moins riches entre les termes. Mais, se serait manquer d'objectivité. En effet, même si on peut considérer les ontologies comme des « descendantes surdouées des thésaurus », leur impact va bien au-delà d'une amélioration de l'accès à l'information. [64 – MENON]

En organisant les domaines de la connaissance, relatifs aux activités de l'entreprise, et en explicitant les liens qui les unissent de manière plus riche et plus nuancée, les taxinomies et les ontologies permettent de fournir une meilleure perception des activités de l'entreprise. Bénéficiaires des technologies de l'Internet, ces « outils » offrent une approche plus intuitive de la recherche d'information. En combinant les fonctionnalités des ontologies et des taxinomies, avec la souplesse de l'hypertexte et le potentiel graphique des interfaces HTML, les chercheurs ont donné naissance à des outils d'exploration beaucoup plus pédagogiques : les Topic Maps. En complète adéquation avec l'univers virtuel et mouvant qu'ils permettent d'explorer, les topic Maps sont en quelque sorte, les premiers outils d'indexation et de recherche qui ne soient pas une simple transposition d'un outil conçu à l'origine pour gérer des documents papier.

Les nombreuses études menées sur la lecture sur écran, ont démontré que les processus intellectuels mobilisés au cours de cette opération étaient très différents de ceux qui sont activés lors d'une lecture traditionnelle sur papier. Dans ce contexte, le cerveau devient plus sensible aux couleurs, aux formes et à l'emplacement des pavés de texte, et il n'analyse plus les informations de la même manière. C'est tout

l'objectif des ergonomes de concevoir des interfaces, qui permettent de traduire des informations textuelles dans un langage graphique mieux adapté à l'univers visuel du web. Et c'est là que réside le point fort des Topic Maps, en adoptant un langage graphique, ces nouveaux outils sont devenus beaucoup plus faciles à manipuler, ce qui permet de rééquilibrer l'accès aux informations entre les utilisateurs néophytes et les experts.

Tenir compte des besoins, mais aussi du niveau de maîtrise des TIC, et de la culture professionnelle de chaque salarié est véritablement le nouvel enjeu des entreprises. Il devient alors nécessaire de créer des espaces intermédiaires qui permettront à l'utilisateur « *de partir de ce qu'il connaît déjà, pour tracer un chemin vers ce qu'il ne connaît pas encore* » [15-CONTAT]. Les taxinomies, en proposant une organisation des connaissances de l'entreprise sous la forme d'une arborescence, permettent aux salariés de s'approprier des points de repères qui leur permettent d'explorer graduellement des zones d'informations vers lesquelles ils ne se seraient pas spontanément dirigés. En personnalisant l'accès aux informations, ils contribuent à améliorer la circulation de l'information dans les entreprises, tout diminuant les clivages entre les corps de métiers. Implantés dans des contextes professionnels hétérogènes, tant du point de vue humain que technique, ils deviennent des outils fédérateurs, qui contribuent à l'amélioration des performances des entreprises.

Comme l'enfant qui commence par apprendre à reconnaître les lettres de l'alphabet avant de les associer pour former des syllabes puis des mots, les systèmes de gestion de contenu ont réintroduit grâce à ses outils, un élément d'une importance capital pour l'avenir de notre société numérique : « le bon sens » !

Ranger, trier, classer, organiser, rien de plus logique pour faciliter l'accès à un document. Si le passage au tout numérique, et l'amélioration grandissante des performances des moteurs de recherche ont pu laisser croire à la disparition programmée des langages documentaires, on s'aperçoit aujourd'hui, que ces langages ont encore un avenir. Liftés, remodelés, les technologies de l'intelligence artificielle leurs ont donné un coup de jeune qui leurs ont permis de retrouver peu à peu leur place dans l'univers numérique, entraînant dans leurs sillages un

repositionnement du métier de documentaliste.

Considérés comme des médiateurs entre les utilisateurs et l'information recherchée, les documentalistes ont depuis toujours été des concepteurs d'outils documentaires destinés à faciliter l'accès aux documents. Or, comme nous l'avons vu dans cette étude, face à l'autonomie croissante des « chercheurs d'information », ces outils, autrefois destinés à l'usage exclusif des documentalistes, sont devenus indispensables au grand public. Par conséquent, leur transposition dans un contexte numérique n'a pu se faire sans provoquer une hybridation du métier de documentaliste. Experts en gestion des fonds documentaires, mais aussi concepteurs d'outils permettant la consultation et la communication de l'information contenue dans ces documents, les spécialistes de l'information ont du intégrer de nouvelles connaissances qui ne faisaient pas partie de leur cœur de métier : langages informatiques, graphisme, ergonomie, linguistique, marketing, sociologie, etc…

C'est en assimilant ces nouvelles techniques que les documentalistes se sont donnés les moyens de jouer un rôle actif dans la réintégration des langages documentaires dans les systèmes de gestion informatisés. Aujourd'hui, la mission des spécialistes de l'information consiste à transférer à ces outils, l'intelligence et les savoirs faire de leur métier, de manière à ce qu'ils occupent la place de médiateur qui était autrefois la leur. Du « front office », les documentalistes sont passés au « back office », sans pour autant dévaloriser leur rôle, bien au contraire. …

Les documentalistes ont toujours eu pour mission de fournir des sources d'informations dont la pertinence résultait de leur capacité à adapter leur sélection au profil du demandeur. Face à un utilisateur qui occupe maintenant une place centrale au sein du système de gestion de l'information, ils sont à nouveau en mesure de se réapproprier une part active dans la conception et l'implantation des logiciels de gestion. Ainsi, ils peuvent agir de manière efficace et concrète sur la perception de l'organisation structurelle de l'entreprise, en proposant des outils fédérateurs qui permettent d'harmoniser l'accès et la circulation des informations entre les salariés des entreprises, tout en tenant compte des particularités de chacun.

Toutefois, il est important de ne pas se laisser éblouir par le « miroir aux alouettes » de la technologie. Sans remettre en cause les performances dont sont capables ces outils, leur implantation dans une entreprise ne transformera pas en un claquement de doigt, les habitudes solidement ancrées des salariés. Face au réflexe « Google », il reste aux documentalistes à « déconditionner » les utilisateurs, afin qu'ils redécouvrent les vertus des classifications.

Bibliographie

Préambule

Cette bibliographie a été arrêtée au 25 septembre 2005. Les références sont classées par thème puis par ordre alphabétique dans chaque thématique.

L'objet de ce mémoire étant d'effectuer un point sur les origines, l'usage et l'avenir des Taxinomies, ontologies, thesaurus et Topic Maps, en étudiant en quoi ils se rattachaient à la famille des langages classificatoires et comment ils contribuaient à améliorer la recherche d'information au sein des entreprises ; j'ai délibérément choisi de privilégier les documents les plus récents (à l'exception des ouvrages généraux abordant la théorie des sciences et techniques de la documentation).

Tous les documents mentionnés dans cette bibliographie ne sont pas cités dans le corps de ce mémoire. Ces derniers ayant plutôt servis à alimenter ma réflexion sur le sujet.

Pour faciliter l'identification des documents cités dans le texte, les références aux ouvrages mentionnent entre crochets le nom de l'auteur et le numéro de la notice, ainsi que le n° de la page pour les citations.

Listes des thématiques :

Information et recherche d'information : notions fondamentales

[1-FROCHOT]

FROCHOT, Didier. *Vous avez dit « traitement de l'information ? »*. [En ligne], mai 2005, [référence du 30/07/2005].

http://www.defidoc.com/info_doc_connaissance/TraitementInfo.htm

[2-CHAUMIER]

CHAUMIER, Jacques. *Des techniques documentaires aux technologies de l'information, du documentaliste au knowledge manager – Quelques réflexions*. [En ligne], novembre 2004, [référence du 30/07/2005].

http://www.defidoc.com/info_doc_connaissance/DesTechDoc.htm

Un article qui évoque les mutations en cours au sein de l'information documentaire, face au développement du document numérique. Il évoque le glissement sémantique qui touche les termes issus de la documentation, lorsqu'ils sont transposés dans l'univers informatique et les nouvelles missions des documentalistes.

[3-BENOIT]

BENOIT, Olivier. *Intranet et gestion des connaissances ... Quand l'informatique « apprend » le bon sens !* . [En ligne], avril 2004, [référence du 30/07/2005].

http://www.defidoc.com/info_doc_connaissance/IntraGestConnaissances.htm

Un article qui remet la révolution informatique en perspective avec les grandes évolutions qui ont touché la civilisation humaine, en rappelant que depuis toujours l'Homme a cherché à organiser son univers et que par conséquent l'âge du net n'échappera pas à ce reflex naturel et nécessaire.

[4-MOLINARO]

MOLINARO, Fabrice. *Document et information*. [En ligne], janvier 2004, [référence du 30/07/2005].

http://www.defidoc.com/info_doc_connaissance/DocumentInfo.htm

[5-ACCART]

ACCART, Jean-Philippe, RETHY, Marie-Pierre. L'utilisateur et la recherche d'information. In : ACCART, J.P, RETHY, M.P, *Le métier de documentaliste*. Paris : Electre – Editions du Cercle de la Librairie, 2003, p.37 –p.61.

Un ouvrage de référence qui propose dans ce chapitre, un rappel des processus intellectuels et techniques qui interviennent lors des différentes étapes de la recherche.

[6-FROCHOT]

FROCHOT, Didier. *Information ou document ?*. [En ligne], décembre 2003, [référence du 30/07/2005].

http://www.defidoc.com/info_doc_connaissance/InfoOuDoc.htm

[7-FROCHOT]

FROCHOT, Didier. *Document, donnée, information, connaissance, savoir*. [En ligne], septembre 2000 – décembre 2003, [référence du 30/07/2005].

http://www.defidoc.com/info_doc_connaissance/DocDonneeInfo.htm

Un article clair et concis qui permet de faire le point sur quelques notions fondamentales des sciences de l'information.

[8-FROCHOT]

FROCHOT, Didier. *Définition subjective de l'information*. [En ligne], septembre 2000 – décembre 2003, [référence du 30/07/2005].

http://www.defidoc.com/info_doc_connaissance/DefSubjInfo.htm

[9-MANIEZ]

MANIEZ, Jacques. *Actualités des langages documentaires*. Paris : ADBS, 2002, 395 p

Un ouvrage de référence incontournable. Il propose une réflexion très complète sur la recherche d'information et les langages documentaires, leur rôle, leur histoire et leur évolution face aux nouvelles technologies. L'auteur évoque aussi la notion de sujet et sa difficile traduction par l'indexation.

[10 – WALLER]

WALLER, Suzanne. *L'analyse documentaire : une approche méthodologique*. Paris, 1999, ADBS

Langages documentaires : historique, théorie et applications

[11-MANUE]

MANUE. *Petite histoire des classifications*. [En ligne], 27 juin 2005, [référence du 24/08/2005].

http://www.figoblog.org/document803.php

[12-DELHERBE]

.DELHERBE, Nicole. *Langages documentaires et thesaurus*. INTD-CNAM - Paris, décembre 2004

Support de cours qui résume les connaissances de bases sur le fonctionnement et les règles de conception des langages documentaires.

[13 – MASSE]

MASSE, Claudine. *Indexation documentaire*. INTD-CNAM - Paris, novembre 2004

[14 – RIVE]

RIVE, Michèle. *Les classifications*. . INTD-CNAM - Paris, novembre 2004

Un bref historique des langages classificatoires, suivi d'une illustration par l'exemple des principales classifications.

[15 -CONTAT]

CONTAT, Odile. *Langages documentaires et nouvelles technologies : l'avenir des langages et leur positionnement au cœur des systèmes d'informations dans le contexte de la presse*. Mémoire de Dess, INTD/CNAM, 2002-2003, 89 p.

Dans ce mémoire l'auteur s'interroge sur la place des langages documentaires traditionnels face aux nouveaux outils d'indexation informatisés. Il étudie la manière dont ils ont évolué et comment ils contribuent au traitement de l'information dans le milieu de la presse.

[16-ACCART]

ACCART, Jean-Philippe, RETHY, Marie-Pierre. Le circuit du document : traitement. In : ACCART, J.P, RETHY, M.P *Le métier de documentaliste*. Paris : Electre- Editions du Cercle de la Librairie, 2003, 451 p

Un inventaire des différents langages documentaires et de leurs applications dans le domaine de la gestion documentaire informatisée.

[17-ROBERGER]

ROBERGER, Olivier. *Convera classifie dynamiquement les résultats*. *01 Informatique*, [en ligne], 07 septembre 2003, [consulté le 22/O0/2005].
www.01net.com/outils/imprimer?article=222217

[18-BLANQUET]

BLANQUET, Marie-France. *Intérêt pédagogique de l'apprentissage des langages documentaires*. [En ligne], janvier 2000, [référence du 16/07/2005].
http://savoirscdi.cndp.fr/culturepro/actualisation/linguistique/blanquet.htm
Dans cet article, l'auteur s'interroge sur le rôle des langages documentaires, en tant qu'outils d'apprentissage de la rigueur intellectuelle pour les élèves. Dans un premier temps il rappelle les propriétés des classifications, puis il analyse leurs apports dans le contexte de la recherche documentaire informatisée.

[19 – AMAR]

AMAR, Muriel. *Les fondements théoriques de l'indexation : une approche linguistique*. Paris : ADBS, 2000, 355 p.
L'auteur expose les caractéristiques de l'indexation, et rappelle les fondements théoriques de cette pratique. Puis il évoque comment les langages d'indexation automatique ont contribués à l'évolution des technologies documentaires.

[20-FEYLER]

FEYLER, Françoise. *De la différence entre langage d'indexation et langage d'interrogation*. [En ligne], mars 1999, [référence du 16/07/2005].

http://savoirscdi.cndp.fr/culturepro/actualisation/bibliotheco/LangageFeyler.htm

En replaçant les langages documentaires dans leur contexte historique, l'auteur retrace leur évolution et la manière dont ils se sont spécialisés au contact des systèmes de gestion informatisée.

[21 – HOLZEM]

HOLZEM, Maryvonne. *Terminologie et documentation : pour une meilleure circulation des savoirs*. [en ligne], 1999, [référence du 16/07/2005].

http://www.accart.nom.fr/Analyses/AnalyseHolzem.html

[22-AFNOR]

AFNOR, *Vocabulaire de la documentation, AFNOR* – Paris-La Défense, 2ème Edition, 1987

Edition en vigueur du glossaire qui fournit les définitions officielles des termes de l'information-documentation.

[23]

UDC CONSORTIUM. [consulté le 16/07/2005].

http://www.udcc.org/cdu.htm

[24 - CHEVILLOTTE]

CHEVILLOTTE, Sylvie. *Les langages documentaires*. [En ligne], [référence du 16/07/2005].

http://repere.enssib.fr/frontOffice/afficheArticle.asp?idTheme=12

[25 - MRAIHI]

MRAIHI, Saida. *Initiation aux outils de recherche sur Internet*. [En ligne], [référence du 09/09/2005].

http://reseauxdoc.u-paris10.fr/s_or/ope_re.htm

Un mode d'emploi des opérateurs de recherche les plus utiles pour formuler une requête dans un système de recherche informatisé.

Outils de recherche et systèmes de gestion de l'information

[26]

ARISEM. [Consulté le 22/09/2005], www.arisem.com

[27]

AUTONOMY. [Consulté le 22/09/2005], www.autonomy.com

[28]

EMC2 DOCUMENTUM. [Consulté le 22/09/2005], www.documentum.com

[29 - CROCHET DAMAIS]

CROCHET DAMAIS, Antoine. *Comprendre la gestion de contenu*. *JDNet Solutions*, [en ligne], 25 août 2004, [référence du 09/09/2005].

http://solutions.journaldunet.com/dossiers/pratique/gestiondecontenu.shtml

Une FAQ sur les notions fondamentales qui touchent à la gestion des contenus numériques.

[30-GUENTHER]

GUENTHER, Kim. *What's your style ? : Organizing Information for the Web*. *OnlineMag*, juillet/août 2004, [référence du 24/07/2005].

http://www.onlinemag.net/

[31 – MOLINARO]

MOLINARO, Fabrice. *Vers la société de l'information – Le défi des nouvelles technologies*. [En ligne], janvier 2004, [référence du 16/07/2005].

http://www.defidoc.com/info_doc_connaissance/SocInfo1DefiNT.htm

L'auteur s'interroge sur le rôle des documentalistes face à l'explosion de l'information numérique.

[32 –SERRES]

SERRES, Alexandre. *Recherche d'information sur Internet, ou en sommes-nous ?, ou allons-nous ?*. [En ligne], 31 août 2004 [référence du 06/O7/2005]. http://savoirscdi.cndp.fr/culturepro/actualisation/linguistique/serres.htm

Un article très complet qui dresse un état des lieux de la recherche sur le web, en analysant les différentes démarches adoptées par les utilisateurs pour satisfaire leur besoin d'information. Puis dans la seconde partie de cet article, l'auteur propose un panorama des outils de recherche. Puis il distingue trois domaines d'évolution possible pour ces outils et détaille les technologies qui leurs sont associées. Il évoque, en particulier toutes les technologies qui permettent d'introduire des relations de type sémantique dans les outils de gestion de contenu.

[33 – INVANCIUC DENIAU]

IVANCIUC DENIAU, Alina. *Moteurs de recherche et restitution de l'information dans les grandes entreprises*. Mémoire de DESS, INTD, 25 Novembre 2003, 154 p.

Une étude très complète sur les moteurs de recherche, leurs différents modes de fonctionnement, et leur impact sur la recherche d'information. L'auteur s'intéresse plus particulièrement au logiciel Verity et explique les fonctionnalités des nouveaux outils issus du TALN : les ontologies, les taxinomies, et les Topic Maps.

[34 – CHARVET]

CHARVET, Prunelle. *Propositions d'outils et de méthodes pour la réalisation d'un portail Internet thématique : le cas du portail des langues de France (www.languesdefrance.com)*. Mémoire de DESS, INTD/CNAM, 1er Octobre 2003, 70 p.

L'auteur dresse une typologie des portails Internet, donne les étapes à suivre pour conduire un projet de ce type, et propose une sélection d'outils et de méthodes pour y arriver.

[35 – COOK]

COOK, Aidan. *The march of the intranet and extranet : what are they ? What can you use them for ? And how can they benefit your organisation ?*. *Managing Information*, juillet/août 2003 - 10:6, p.33-36

[36- FROCHOT]

FROCHOT, Didier. *Les moteurs de recherche*. [En ligne], décembre 2000 – décembre 2003, [référence du 21/07/2005].
http://www.defidoc.com/Rech_info_int/Moteurs.htm

[37- FROCHOT]

FROCHOT, Didier. *Les répertoires ou annuaires*. [En ligne], décembre 2000 – décembre 2003, [référence du 24/07/2005].
http://www.defidoc.com/Rech_info_int/Repertoires.htm

[38 – GRIVEL]

GRIVEL, Luc. *Panorama des méthodes et des technologies pour l'accès et la restitution des documents*. [En ligne], [référence du 09/09/2005].
http://www.gfii.asso.fr/sem4_paris1.htm
L'auteur analyse les étapes successives du traitement des documents (indexation, extraction de l'information, et classification des concepts) par les moteurs de recherche.

[39 - VILLANOVA-OLIVIER]

VILLANOVA-OLIVER, Marlène. *Adaptabilité dans les systèmes d'information sur le Web : Modélisation et mise en œuvre de l'accès progressif*. Thèse – Spécialité Informatique, Institut national polytechnique de Grenoble, 18 décembre 2002, p.1-24

Une étude technique des propriétés des systèmes d'information du web, où l'auteur aborde l'intérêt des technologies hypertextes pour la mise en place de systèmes d'accès progressif à l'information pertinente. Il aborde également les technologies permettant de concevoir des modèles utilisateurs susceptibles de fournir un accès personnalisé à l'information dans l'univers du web.

[40-REMIZE]

REMIZE**,** Michel. *Sites d'entreprise : La dispersion appelle la reconstruction*. *Archimag*, mai 2002 n°154, p.25-28

[41-REMIZE]

REMIZE, MICHEL, *Caroline WIEGANDT : « Nous passons d'un traitement documentaire à un travail sur les contenus et leurs moyens d'accès »*. A*rchimag*, mai 2002 n°154, p.17-18

[42-AZAR-EXBRAYAT]

AZAR-EXBRAYAT, Sophie. *Un intranet documentaire au service de l'usager : conception et mise en place*. *Documentaliste – Sciences de l'information*, 2002, vol. 39, n°4-5, p.190-201

[43 – FERCHAUD]

FERCHAUD, Bernadette. *Comment mettre en œuvre et alimenter un portail d'entreprise*. *Documentaliste – Sciences de l'information*, 2002, vol. 39, n°4-5, p.232-233

Après avoir listé les différentes étapes qui rythment la conception d'un portail d'entreprise, l'auteur rappelle l'importance qui doit être accordée au choix du moteur de recherche et à la sélection des futurs contenus. Il termine son exposé par quelques exemples de portails.

[44-FERCHAUD]

FERCHAUD, Bernadette. *Intranet : conception, réalisation, usages*. *Documentaliste – Sciences de l' information*, 2001, vol. 38, n°3-4, p.220-221

L'auteur dresse un inventaire des fonctionnalités des réseaux intranet et évoque leur impact sur le fonctionnement des entreprises. Puis il évoque les défauts de cet outil et les solutions qui sont envisagées pour y remédier.

[45-CROCHET-DAMAIS]

CROCHET DAMAIS, Antoine. *Moteurs de recherche : le tableau des solutions*. *JDNet Solutions*, [en ligne], août 2001, [référence du 09/09/2005].
http://solutions.journaldunet.com/printer/010829_panoramamoteur.shtml

[46-ALERI]

ALERI, Frédéric ; LAFONT, Denis ; MACARY, J-F. *Le projet intranet. De l'analyse des besoins de l'entreprise à la mise en œuvre des solutions*. Paris : Eyrolles, 1998 (2ème édition), ISBN 2-212-09038-2-.

Un ouvrage de référence qui analyse en détail toutes les étapes d'un projet intranet, en les mettant en perspective avec la structure et l'évolution des entreprises à l'aube du XXIème siècle. Plusieurs chapitres sont consacrés aux moteurs de recherche et aux langages d'indexation et de structuration des documents numériques (XML, SGML, etc. ..)

ONTOLOGIES, TAXINOMIES, THESAURUS, ET TOPIC MAPS

[47-BORDAGE]

BORDAGE, Frédéric. *Les bases du web sémantique*. *01net.com*, [en ligne], août 2005, [référence du 22/09/2005].
http://www.01net.com/outils/imprimer.php?article=239827

[48-TEXIER]

TEXIER, R.. *Taxinomies, thésaurus et ontologies*. *EliKya, intelligence des organisations*, 23 juin 2005, p.1-3

[49-PLOSKER]

PLOSKER, George. *The information strategist. Taxonomies : facts and opportunities for information professionals*. *Online*, janvier/février 2005, p.58-60

L'auteur commence par rappeler que les taxinomies sont avant tout des classifications. Après avoir fait l'inventaire de leurs propriétés, il poursuit son étude en s'intéressant à leur rôle au sein des entreprises et termine en analysant comment ces outils s'intègrent dans la technologie des moteurs de recherche.

[50-SCHEIDER]

SCHNEIDER, Oliver. *Discuter : taxinomie*. [en ligne], septembre 2004, [référence du 07/08/2005].
http://fr.wikipedia.org/wiki/Discuter:Taxinomie

[51 – REGUER]

REGUER, David Emmanuel. *Verity rend dynamique la création des taxonomies*. *01 Réseax*, [en ligne], 01 septembre 2004, [consulté le 22/09/2005].
www.01net.com/outils/imprimer.php?article=256648

Bref aperçu du processus de création des taxinomies dans Verity

[52-JAVED]

JAVED, Mostafa. *Document search interface design : background and introduction to special topic section*. *Journal of the american society for information science and technology*, 55 (10), 2004, p.869-872.

[53-GARSHOL]

GARSHOL, Lars Marius. *Metadata ? Thesauri ? Taxonomies ? Topic maps ! Making sense of it all*. *Journal of Information*, 30 avril 2004, p.378-391

Dans cet article, l'auteur expose qu'elles sont les fonctionnalités de ces outils et comment ils peuvent contribuer à améliorer la pertinence des recherches et de l'indexation automatique, ainsi que l'accès « physique » aux documents.

[54 – BORDAGE]

BORDAGE, Frédéric. *Le web sémantique automatise la gestion du savoir*. *01 Informatique*,[en ligne], 15 avril 2004, [consulté le 22/09/2005].
www.01net.com/outils/imprimer.php?article=276948

Cet article traite de la fusion de Mondeca et Temis, et donne une rapide description des fonctionnalités de leur produit

[55- MORRIS]

MORRIS, Jeff. *Putting it together : taxonomy, classification, & search*. *Transform magazine*, [en ligne], septembre 2003, [référence du 08/08/2005].
http://www.transformmag.com/db_area/archs/200309/tfm0309f2_1.shtml

[56-BERTOLUCCI]

BERTOLUCCI, Katherine. *Happiness is taxonomy : four structures for Snoopy*. *Information Outlook*, mars 2003, p.36-44

[57-JERNST]

JERNST. *What are the differences between a vocabulary, a taxonomy, a thesaurus, an ontology, and a meta-model ?*. [en ligne], 15 janvier 2003, [référence du 08/08/2005].

http://www.metamodel.com/article.php?story=20030115211223271&mode=print

[58-GILCHRIST]

GILCHRIST, Alan. *Thesauri, taxonomies and ontologies – an etymological note*. *Journal of Documentation*, Vol. 59 No. 1, 2003, p.7-18

L'auteur commence par rappeler le sens premier de chacun de ces termes, puis il étudie comment et pourquoi ils ont été « récupérés » pour désigner des outils mis au point par les spécialistes de l'intelligence Artificiel. Puis, il détaille pour chaque terme, les différents usages qu'il a rencontrés. Il dresse en particulier un inventaire très complet des fonctionnalités associées aux taxinomies. Puis explique en quoi ils se rattachent aux langages traditionnels et comment ils ont évolué pour s'adapter à l'univers du web.

[59-CHUANG]

CHUANG, Shui-Lung ; CHIEN, Lee-Feng. *Automatic query taxonomy generation for information retrieval applications*. *Online Information review*, 2003 Volume 27 no. 4, ISSN 1468-4527, p.243-255

[60-ADAMS]

ADAMS, Katherine. *The semantic web*. *OnlineMag*, [en ligne], Juillet/août 2002, p.20-23.

http://www.onlinemag.net/

[61-BAMBURY]

BAMBURY, Paul. *Taxonomy*. *Managing Information*, mai 2002 No. 9, p.28-30

[62-AINSBURY]

AINSBURY, Bob. *Cataloging's comeback*. *Online*, mars/avril 2002, p.27-31

Dans le contexte des portails d'entreprise, l'auteur rappelle les principes de la recherche d'information et explique comment les taxinomies peuvent devenir un outil de pertinence au service des employés de l'entreprise.

[63-GILCHRIST]

GILCHRIST, Alan. *Corporate taxonomies : report on a survey of current practice*. *Online Information review*, 2001 Volume 25 No. 2, p.94-102

Cette étude analyse les processus de construction des taxonomies d'entreprises, et décrit leurs différentes fonctionnalités. Il s'intéresse aux causes de l'ambiguïté attachée à ce terme et recense les relations et les différences qui existent entre les classifications, les thésaurus et les taxinomies. Il illustre son propos en relatant 6 des 22 études de cas, qui lui ont permis d'établir une typologie des différents contextes d'utilisation des taxinomies.

[64-MENON]

MENON, Bruno. L'évolution des langages documentaires. [en ligne], [référence du 08/09/2005].
http://www.bmenon.net/Evolution des langages documentaires.htm

Dans cet article, l'auteur propose une généalogie des langages documentaires, et insiste sur les facteurs qui les ont fait évoluer au cours du XXème siècle.

[64*]

WIKIPEDIA , *Topic Maps,* [en ligne], [consulté le 20/08/2005]
http://fr.wikipedia.org/wiki/Topic_map).

LINGUISTIQUE ET TALN

[65 – MOLINARO]

MOLINARO, Fabrice. *Le patrimoine numérique : introduction*. [en ligne], avril 2005, [référence du 29/08/2005].

http://www.defidoc.com/info_doc_connaissance/PatrimoineNumeriqueIntro.htm

[65 – GRILLL]

GROUPE DE REFLEXION SUR LES INDUSTRIES DE L'INFORMATION ET LES INDUSTRIES DE LA LANGUE (GRIIIL).

Le traitement automatique des langues dans les industries de l'information. Paris, GRILLL, janvier 2005

Une étude récente et très complète de l'industrie du traitement automatique du langage, qui propose dans une première partie un panorama du marché industriel du TAL. Ensuite, les auteurs procèdent à un bref rappel des étapes techniques du traitement linguistique, puis ils proposent une analyse de 6 contextes d'application.

[66-LESPINASSE-SABOURAULT]

LESPINASSE-SABOURAULT, Karine. *Introduction à la linguistique : son application à la documentation : les dictionnaires électroniques*. Paris, novembre 2004

[67-BONNET]

BONNET, Bernard. *De SGML à XML*. Paris, novembre 2004

[68 -GRIMES]

GRIMES, Seth. *The word on text mining*. [en ligne], 10 décembre 2003, [référence du 16/08/2005].

http://www.intelligententerprise.com//031210/619decision1_1.jhtml

[69 – FROCHOT]

FROCHOT, Didier. *Informatique documentaire*. [en ligne], octobre 2003, [référence du 29/08/2005].

http://www.defidoc.com/infor_doc/InformatiDoc.htm

[70- CROCHET-DAMAIS]

CROCHET-DAMAIS, Antoine. *L'automatisation du traitement des contenus passe par le Web sémantique*. *JDNet Solutions*, [En ligne], 24 septembre 2003, [référence du 09/09/2005].

http://solutions.journaldunet.com/0309/030924_websemantic.shtml

[71- CROCHET-DAMAIS]

CROCHET-DAMAIS, Antoine. *OWL : naissance d'un nouvel outil sur le terrain du Web sémantique*. *JDNet Solutions*, [En ligne], 22 août 2003, [référence du 09/09/2005].

http://solutions.journaldunet.com/0308/030822_owl.shtml

[72-BORDERIE]

BORDERIE, Xavier. *Le web sémantique*. *JDNet Solutions*, [En ligne], 18 novembre 2002, [référence du 09/09/2005].

http://developpeur.journaldunet.com/tutoriel/xml/021115xml_websemantique1b.shtml

[73-DALBIN]

DALBIN, Sylvie ; SALLERAS, Bruno. *Une expérience d'utilisation d'un système d'information documentaire en langage naturel*. *Documentaliste – Sciences de l'Information*, 2000, vol. 37, No. 5-6, p.312-324

En s'appuyant sur l'étude du centre documentaire des AGF, l'auteur analyse l'impact de l'installation d'un système de gestion de contenu en langage naturel sur l'organisation du service. Après avoir déterminé les profils des différents utilisateurs, il étudie leurs changements d'habitudes, et analyse la pertinence des réponses obtenues à une série de questions conçues pour tester les limites du dispositif.

[74]

WIKIPEDIA, [consulté le 16/07/05]

http://fr.wikipedia.org/wiki/TALN

[74*]

LAZINIER, Emmanuel ; *XML expliqué au débutant* . [en ligne], 21 décembre 1999, [consulté le 10/08/2005] http://www.chez.com/xml/initiation

Notions d'ergonomie

[75-BERTIN]

BERTIN, Karin. *L'ergonomie des sites web*. *Documentaliste – Sciences de l'Information*, 2005 vol. 42 No. 1, p.58-61

[76-CNRS]

CNRS. *Règles d'ergonomie*. [En ligne], [référence du 07/07/2005]. http://www.dsi/cnrs.fr/conduite-projet/phasedeveloppement/technique/basdevtech5.htm

[77-CASANOVA]

CASANOVA, Xavier ; COHEN, Joëlle. *L'écran efficace : une approche cognitive des objets graphiques*. *Documentaliste – Sciences de l'Information*, 2001 vol. 38 No. 5-6, p.272-289

Les auteurs s'intéressent à l'émergence d'une nouvelle culture visuelle, et s'interrogent sur les apports de la psychologie cognitive pour concevoir des outils documentaires mieux adaptés à l'univers du Web.

Management de l'information dans les entreprises

[78-DELSOL]

DELSOL, Emmanuelle ; BURGER, Clarisse ; BLANC, Sylvie. *La difficile analyse du corpus de connaissances*. [En ligne], 6 mai 2005, [référence du 22/09/2005]. http://www.01net.com/outils/imprimer.php?article=278360

[79-GUYOT]

GUYOT, Brigitte. *Information et organisations : Comprendre les implications d'un management de l'information*. Paris, janvier 2005

[80-BATTISTI]

BATTISTI, Michel. *Un métier, des métiers : convergences et spécificités des métiers des archives, des bibliothèques et de la documentation*. *Documentaliste – Sciences de l'Information*, 2005, vol. 42 No. 1, p.48-57

[81-BORREL]

BORREL, Guillemette. *Documentaliste/Informaticien : dualisme ou partition en duo*. DESS, INTD, 6 décembre 2004, 86 p.

[82-MOLINARO]

MOLINARO, Fabrice. *Vers la société d'information (II) Un rôle stratégique au sein de l'entreprise*. [En ligne], janvier 2004, [référence du 07/07/2005] http://www.defidoc.com/info_doc_connaissance/SocInfo2RoleStrat.htm

Après avoir évoqué l'importance stratégique de l'information dans les entreprises, l'auteur s'intéresse au rôle et l'avenir des spécialistes de l'information au cœur des organisations.

[83-FROCHOT]

FROCHOT, Didier. *La fonction documentaire dans les organisations*. [En ligne], Décembre 2003, [référence du 07/07/2005]
http://www.defidoc.com/info_doc_connaissance/FonctionDocOrg.htm

[84-ACCART]

ACCART, Jean-Philippe. *L'information profilée*. *Archimag*, juin 2002 No. 155

[85-MANIEZ]

MANIEZ, Jacques ; MUSTAFA EL HADI, Widad. *Organisation des connaissances en vue de leur intégration dans les systèmes de représentation et de recherche d'information*. Lille : UL3 – travaux et recherche, 1999, 398 p.

[86-SUTTER]

SUTTER, Eric. *Apprendre à caractériser les cultures professionnelles*. *Documentaliste – Sciences de l'Information*, 1996 vol. 33 No. 2, p.86-88

Documents Air France

[87-BLANCHARD]

BLANCHARD, Audrey. *2005 Annual report Verity Air France*. Roissy, 11 octobre 2005

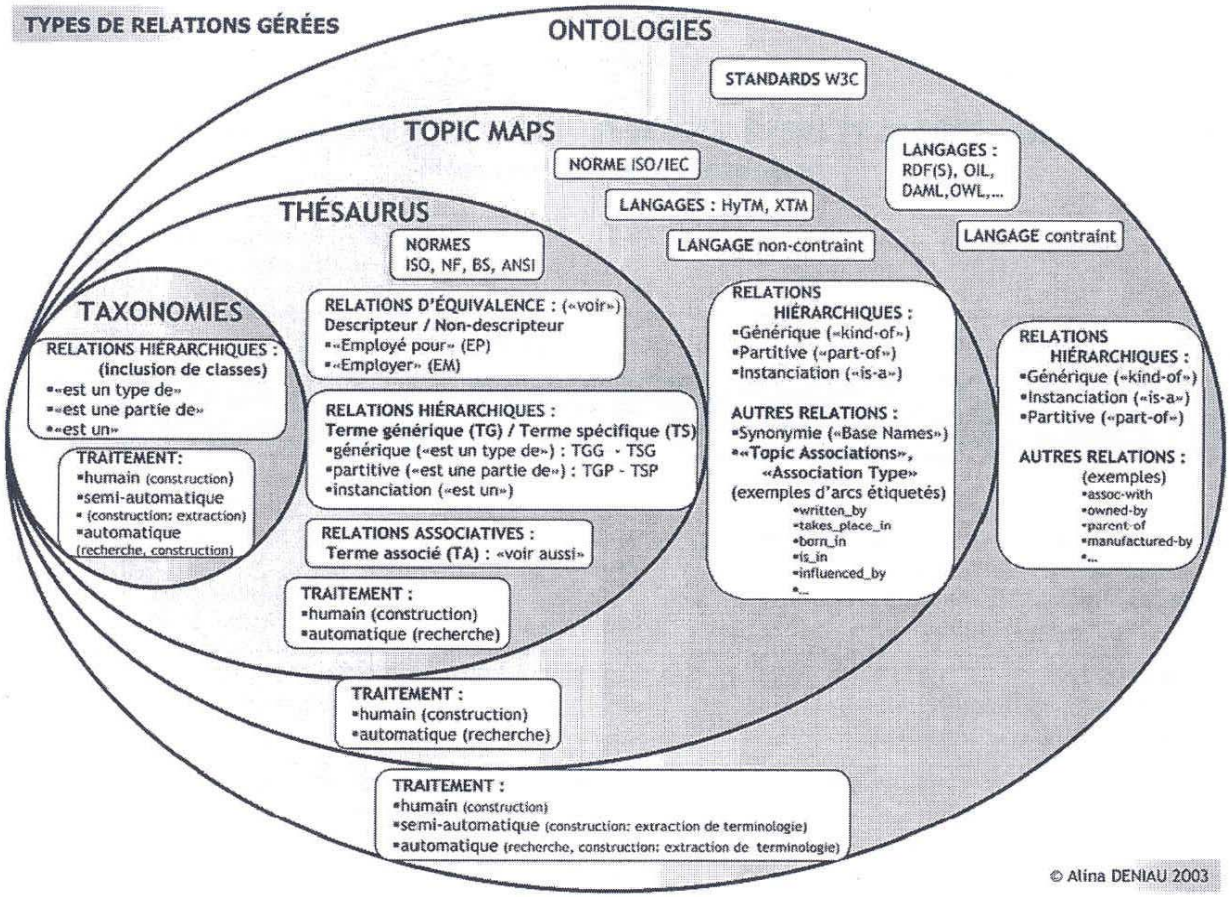
[88 – BLANCHARD]

BLANCHARD, AUDREY . Suivi de gestion des connaissances – Programme intranet de la compagnie Air France, rapport de stage, DESS Innov'Info, 2003

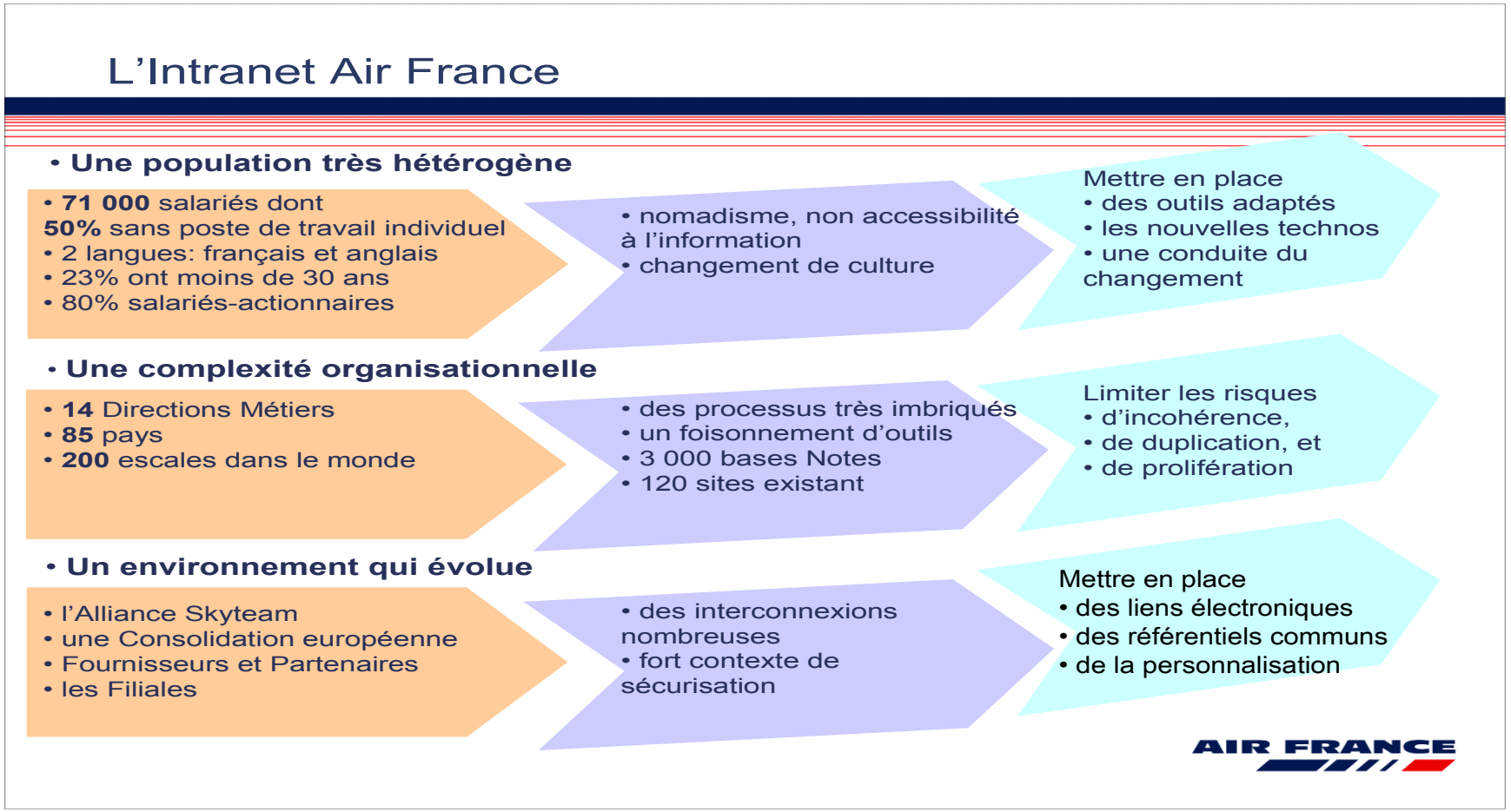
Annexes

Annexe 1 : Représentation graphique des types de relations gérées par les systèmes d'organisation des connaissances (KOS)

Représentation schématique des types de relations gérées par les systèmes d'organisation des connaissances (KOS)



Annexe 2 : Présentation de l'Intranet AF par Audrey Blanchard pour le Forum annuel Verity


L'Intranet Air France
• Une population très hétérogène
• 71 000 salariés dont 50% sans poste de travail individuel
• 2 langues: français et anglais
• 23% ont moins de 30 ans
• 80% salariés-actionnaires
• nomadisme, non accessibilité à l'information
• changement de culture
Mettre en place
• des outils adaptés
• les nouvelles technos
• une conduite du changement
• Une complexité organisationnelle
• 14 Directions Métiers
• 85 pays
• 200 escales dans le monde
• des processus très imbriqués
• un foisonnement d'outils
• 3 000 bases Notes
• 120 sites existant
Limiter les risques
• d'incohérence,
• de duplication, et
• de prolifération
• Un environnement qui évolue
• l'Alliance Skyteam
• une Consolidation européenne
• Fournisseurs et Partenaires
• les Filiales
• des interconnexions nombreuses
• fort contexte de sécurisation
Mettre en place
• des liens électroniques
• des référentiels communs
• de la personnalisation
AIR FRANCE

Annexe 3 : les acteurs du projet Intranet – Audrey Blanchard- rapport de stage innovinfo 2003

Ce document étant confidentiel, il n'a été communiqué qu'aux membres du jury de soutenance.

Annexe 4 : Inventaire des sources d'informations du service Sécurité et conditions de Travail

INVENTAIRE DES RESSOURCES DOCUMENTAIRES DU SERVICE SANTE SECURITE DU TRAVAIL

FONCTION OCCUPEE	INFORMATION PRODUITE	SOURCES D'INFORMATIONS INTERNES	SOURCES D'INFORMATIONS EXTERNES	SOURCES D'INFORMATIONS EN ATTENTE
Préventeurs	Fiches techniques EPI. Documents pour formation réseau élargi. Procédures générales Précis méthodologique. Réalisation de dossiers documentaires par type de risques. "bibliothèque de sites web spécialisés"	Secur-e-flash / Secur-e-doc. Questions directes au chargé de veille. Documents divers (y compris vidéos) de la médiathèque Spad Base alerte	Cédérom Editions législatives. Sites Internet : INRS / LEGIFRANCE / EUROLEX BTS / OIT / AINT67 / BOSSONS FUTE MINISTERE DE L'EMPLOI / CRAM / CRAMIF SITE DE L'AGENCE NATIONALE POUR L'AMELIORATION DES CONDITIONS DE TRAVAIL Participation à des colloques et manifestations SST	Versions numériques des **Editions "Liaisons Sociales"** : Santé, sécurité & conditions de travail au quotidien. Fiches pratiques TISSOT. Guide Santé, sécurité TISSOT.
Ergonomes	Document de sensibilisation à l'ergonomie (destiné aux postes d'encadrement). Cahier des charges. Procédure générale. Précis méthodologique. Présentation ergonomie (stage préventeur)	Intralignes A.F : portails métier. Secur-e-flash / Secur-e-doc Questions directes au chargé de veille. Annuaire hiérarchique Lotus. Documents médiathèque.	SITES INTERNET : revue@ctivité-piste Actes des congrès en ligne et liste de diffusion du SELF INRS, ANACT, ERGOLAB. ergo.list (bientôt payant). JOURNAUX & OUVRAGES DE REFERENCE. Travail & Sécurité. Bulletin de liaison SELF. Ouvrages d'ergonomie (pas de version numérique). Participation à colloques	REVUES SPECIFIQUES ENTITES : Briefing Escale Point fixe (Récupérer sous format papier uniquement pour l'instant)
Chargé de veille réglementaire	Bulletins mensuels et trimestriels d'information juridique et technique: Secur-e-flash. Secur-e-doc. Réponses personnalisées aux questions juridiques et/ou techniques à l'attention du réseau élargi. Procédure générale. Dossiers thématiques.	portails métiers sur Intralignes. Médiathèque. Isotarget. Observations / contacts sur le terrain et avec le réseau	Revues juridiques (ex :droit social) et techniques (ex : Travail & Sécurité). Participation à des groupes d'entreprises / s Séminaires. Actes de colloques. Sites Internet spécialisés (idem préventeurs).	A l'attention du réseau, 10 accès simultanés aux versions numériques des : Editions "Liaisons Sociales" : Santé, sécurité & conditions de travail au quotidien. Editions Législatives" : Sécurité et conditions de travail

Chargé de communication	Affiches. Agenda sécurité. Annuaire du réseau sécurité	Portails métiers sur Intralignes. Médiathèque. Isotarget. Observations / contacts sur le terrain et avec le réseau. Echanges avec les préventeurs.	Entreprises externes. Sites Internet spécialisés en communication	
Correspondants sécurité	Flash sécurité. Rapports d'accidents de travail. Analyses statistique. Supports manifestations sécurité. Rédaction de plans de prévention	Secur-e-flash / Secur-e-doc. Questions directes au chargé de veille. ISOTARGET. SPAD. Base alerte	site Legifrance. Sites sécurité du travail. Sites des Editions législatives. Sites des ministères	Accès numérique aux Editions législatives et liaisons sociales. Amélioration de l'accès à la base Secur-e-doc. (distribution liens directs en cours).
Médecins	Résultats en attente	Résultats en attente	Résultats en attente	Résultats en attente
CHSCT	Compte rendu de réunion. Compte -rendu d'inspection. Tableau de suivi des accidents du travail. Enquête accident. Avis motivé sur question sécurité.	membres du réseau. Personnel de l'établissement. Préventeurs. Chargé de veille réglementaire. Médecins du travail. Informations échangées au cours des réunions. Secur-e-doc/ Secur-e-flash ISOTARGET. SPAD. Base alerte		Accès numérique aux Editions législatives et liaisons sociales. Amélioration de l'interface Secur-e-doc. (portail Intranet centralisé et accès aux fonctionnalités du moteur de recherche Verity). Consultation des nouveaux documents sous formats PDF.