CONSERVATOIRE NATIONAL DES ARTS ET MÉTIERS

INSTITUT NATIONAL DES TECHNIQUES DE LA DOCUMENTATION

MÉMOIRE PRÉSENTÉ EN VUE D'OBTENIR

LE DESS EN SCIENCES DE L'INFORMATION
ET DE LA DOCUMENATION SPÉCIALISÉES

par Odile Contat

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 soutenu devant un jury, composé de :
Mme Claudine Masse
Mme Annie Milhaud

Année 2002-2003
CYCLE SUPERIEUR PROMOTION XXXIII

Mes remerciements à Claudine Masse et Annie Milhaud pour leur soutien et leurs conseils.
Je remercie également les personnes qui ont bien voulu me recevoir et répondre à mes questions au journal Le Monde, Les Échos et au Nouvel Observateur.

PLAN DU MÉMOIRE

Annexes

Notice du mémoire

INTRODUCTION

« Toute recherche d'information textuelle part du langage et aboutit à des produits du langage » [2-Maniez, p°7]

Index, glossaire, plan de classement, classification, thésaurus ; Depuis toujours, la nécessité de retrouver l'information contenue dans les documents a engendré la création de langages particuliers.
L'objectif est toujours le même : décrire le contenu d'un document, définir son sujet, ses caractéristiques afin de pouvoir plus facilement et plus rapidement retrouver l'information désirée lors d'une recherche.
Les langages documentaires ont répondu à la nécessité de se doter d'un instrument de contrôle linguistique pour maîtriser la richesse, l'ambiguïté et la polysémie du langage naturel, pour expliciter un sujet, que ce soit lors de la description d'un document ou lors de la recherche d'informations.

Ces langages ont pris bien des formes depuis les premiers index matière des encyclopédies jusqu'aux thésaurus intégrés dans les bases de données.
Ils ont été pendant de nombreuses années les garants de l'accès à l'information.
Dans les bibliothèques ou les centres de documentation, chacun était familiarisé à l'utilisation des fiches cartonnées et à la recherche dans les énormes catalogues matière ou auteur.

Puis peu à peu les techniques informatiques, la numérisation des documents et les nouveaux outils de recherche ont transformé les modes d'accès à l'information.

Dans un premier temps, les logiciels de base de données ont permis l'informatisation des catalogues, remplaçant les fiches cartonnées par des notices informatiques. La recherche des références s'effectuait sur la base de données, mais la consultation des documents perdurait sous format papier.
L'existence de ces langages documentaires et de leurs fonctions, des plans de classement et de l'indexation humaine n'étaient pas remis en cause. Au contraire, l'informatique a permis de mieux gérer ces langages et de les intégrer dans des outils d'aide à l'interrogation et à la recherche.

Puis un double phénomène transforma profondément la gestion des documents et la recherche d'informations.
D'une part la numérisation a permis la consultation du document en texte intégral, sous format informatique, et plus seulement l'accès à une notice le décrivant.
D'autre part, le développement des logiciels permettant l'indexation automatique et la recherche sur le texte intégral, a bouleversé les perspectives de la recherche d'informations, et a positionné ces nouveaux outils comme des concurrents directs des langages documentaires.
Pourtant, beaucoup de systèmes ont continué à utiliser les langages documentaires en cumulant, dans les bases de données, indexation humaine et recherche par mots-clés et indexation automatique et recherche directe sur le texte.

Aujourd'hui, l'essor des systèmes de recherche sophistiqués, alliant les performances des traitements linguistiques et sémantiques à la recherche automatisée, les avancées en matière d'intelligence artificielle et de bases de connaissances, posent avec plus d'acuité encore la problématique de l'utilité des langages documentaires.

Dès lors, on est en droit de se demander si ces nouvelles technologies remettent en cause l'utilité, la nécessité et l'intérêt de tout langage documentaire que ce soit lors de la description d'un document ou lors de la recherche d'informations.

Il ne s'agit pas ici d'analyser les formes ou les structures du langage documentaire mais ses fonctions par rapport à la recherche d'informations.
Il ne s'agit pas seulement de décrire ou d'évaluer les performances des moteurs de recherche mais de comprendre comment l'évolution technologique a mis peu à peu en cause les langages documentaires.
Il s'agit de tenter de voir comment ces langages peuvent avoir encore un rôle et une place face aux nouvelles technologies notamment linguistiques et sémantiques.

J'ai choisi de me limiter dans ce questionnement. D'abord, en ne prenant en compte que les problèmes liés à la recherche sur des documents textuels, les problématiques de l'image étant trop différentes. Ensuite, en me concentrant sur le contexte de la presse que j'ai pu observer lors de mon stage, et qui est très représentatif des difficultés de l'analyse de contenu, de l'ambiguïté du langage naturel et des problématiques de la recherche d'informations.

Dans une première partie je définirai précisément les fonctions du langage documentaire et les spécificités des nouvelles technologies. Puis, je parlerai plus précisément de la problématique de la recherche d'informations et du contexte de la presse, avec mes sources sur le terrain à la fois chez Bayard et lors de mes entretiens. Ensuite, en partant des besoins d'informations des journalistes, je tenterai de définir les rôles et les fonctions des langages documentaires face à l'utilisation des nouvelles technologies. Enfin j'essaierai de tirer des conclusions sur l'avenir de ces langages par rapport aux nouvelles technologies.

I - DEFINITIONS

La recherche de documents est fondée sur le rapprochement entre les termes d'une requête et les termes décrivant le contenu d'un document.
Pour retrouver un document, pour rechercher une information, il faut donc avoir au préalable décrit ce document et caractérisé les informations qu'il contient, c'est à dire indexé.

Indexer dit le Robert c'est attribuer à un document une marque distinctive renseignant sur son contenu et permettant de le retrouver.
« C'est un processus qui a pour objectif de représenter le contenu d'un document sous une forme condensée (un ensemble de termes), c'est-à-dire passer d'un document à sa représentation manipulée par un système documentaire. » [11-Le Loarer, p°150]

Que la recherche s'effectue dans un fonds papier ou dans une base de données, le mécanisme de l'indexation est indissociable de celui de la recherche d'informations.
On peut distinguer deux grands types d'analyse et de description des documents :

- l'indexation humaine par un langage documentaire
- l'indexation automatisée du texte intégral

La première approche se place du point de vue documentaire :
« L'indexation est une opération intellectuelle impliquant une analyse approfondie d'un document et la représentation condensée de l'information portée par ce document. » [12-Chaumier et Dejean, p°14]

La seconde approche se situe du point de vue informatique :
« L'informatique considère l'indexation comme un repérage des informations dans un ensemble de documents, opération qui permet d'accélérer le processus de la recherche de l'information. » [12-Chaumier et Dejean, p°15]

Dans le premier cas, l'indexation et la recherche passent par la médiation du langage documentaire. Dans le second cas, ce sont des outils informatiques et linguistiques qui sont au cœur du traitement des documents et de la recherche de l'information.

A - Les langages documentaires : définition et fonctionnalités

1 - Les langages d'indexation et de recherche : définition et historique

Langage documentaire, langage d'indexation, langage contrôlé, langage artificiel… de quoi parle–t-on ?
Reprenons la définition de l'Afnor pour y voir plus clair :
« Les langages documentaires sont des langages artificiels fait de représentation 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 » [2-Maniez, p°206]

Les langages documentaires sont donc des passerelles, des traducteurs entre le langage de l'auteur et celui du demandeur ; ils sont dits contrôlés parce que les termes qui les composent doivent être utilisés à l'identique lors de l'indexation documentaire et lors de la recherche ; ils sont dits artificiels car ils sont construits pour formaliser de façon synthétique le contenu des textes c'est-à-dire pour traduire les sujets, aspects, thèmes, etc. d'un document par des concepts clairs et sans ambiguïté.

Une autre définition, plus récente, résume bien les fonctions des langages documentaires :
« Les langages documentaires sont des codes sémantiques de représentation des sujets, permettant à un système documentaire de repérer les documents par une formulation rigoureuse de leur contenu, et aux utilisateurs d'ajuster leurs interrogations à ces formulations » [2-Maniez p° 207]

Les premiers index matière à la fin des ouvrages correspondaient déjà à cette problématique d'analyse du contenu et de définition du sujet.

Puis, dans les bibliothèques, au XIX siècle, sont apparues les classifications universelles (Dewey, CDU, Bibliothèque du Congrès).
Ces langages organisent les connaissances en un système ordonné de classes et de sous classes hiérarchisées. A chaque branche du savoir correspond une classe et des sous classes qui sont elles-mêmes subdivisées pour prendre en compte tous les sujets.
Cette organisation hiérarchique du savoir permet à la fois le regroupement intellectuel par sujet, et le classement physique des documents puisque à chaque sujet correspond un code, un indice de classement.

Puis au milieu de notre siècle, des langages dits analytiques ont été construits afin de décrire le sujet non plus globalement mais par une combinaison de concepts.
Les plus simples de ces langages sont les listes d'autorité qui définissent un terme unique pour chaque concept : chaque terme est normalisé (féminin, pluriel, etc.), clarifié (définition) et lié à une liste de synonymes pour lesquels il sera employé (liste de renvois).

Le principe est le même dans les thésaurus, mais les descripteurs sont structurés en catégories et reliés entre eux par des relations sémantiques : la relation hiérarchique (est le père de, est le fils de) et la relation associative (terme proche sémantiquement : voir aussi).

Il existe bien d'autres formes de langages documentaires comme les classifications à facettes ou le langage d'indexation matière Rameau adopté par la bibliothèque nationale.
Mais quelles que soient les formes prises par les langages documentaires, ce qui importe ce sont leurs fonctionnalités par rapport à la recherche d'informations.

2 - Fonction de normalisation et de désambiguïsation du langage naturel

Tous les langages documentaires ont été élaborés afin d'éliminer les problèmes d'ambiguïté du langage naturel, et de fournir une représentation univoque et formalisée des documents.

Si l'indexation et la recherche étaient formulées en langage naturel, c'est-à-dire dans une langue usuelle sans aucune contrainte, la formulation de la question par l'utilisateur serait forcément différente de celle du sujet par l'indexeur. Cela provoquerait donc des difficultés insurmontables à la recherche.
Un texte, indexé par « nocivité du tabac », ne serait pas retrouvé pour une question sur « le danger des cigarettes ».

L'utilisation d'un langage documentaire normalisé permet d'ajuster la question à la formulation du sujet sans confusion possible.
Les règles d'écritures présentes dans tous les langages documentaires normalisent les termes, notamment les noms propres, et lèvent les difficultés liées aux multiples orthographes possibles.

Chaque mot-clé ou descripteur est normalisé à sa forme la plus simple, généralement un nom commun masculin et singulier, auquel seront ramenées toutes les formes dérivées pluriel, féminin, etc.

Définir le sens exact de chaque terme d'un langage permet d'éviter les nombreux pièges du langage naturel : polysémie, synonymie, homonymie, etc. *(Voir annexe 1, p°I)*
Ainsi, tous les concepts d'une classification font partie d'une chaîne hiérarchique, ce qui offre une solution aux problèmes de l'homonymie.
Par exemple, le concept de calcul perd son ambiguïté lorsqu'il reçoit l'indice 517 ou 616. Dans le premier cas, il s'agit de calcul mathématique, dans le second, de calcul médical.

Un terme polysémique tel que « décoration » sera clairement définie dans un thésaurus comme étant employé dans le sens « récompense, médaille » et pas dans celui « d'agencement intérieur d'une maison ».

De même, on réduit les problèmes de synonymie en établissant des relations d'équivalence, en choisissant au préalable le descripteur automobile plutôt que voiture.

Les langages documentaires codifient l'information. Cette codification unique permet de rassembler des documents sous le même terme ou concept même si les auteurs n'utilisent pas le même vocabulaire pour parler d'une notion, même si ce vocabulaire évolue.

L'indexation humaine tient compte de l'implicite et des non-dits des textes et de la connotation des mots.

Le sens et la forme de chacun des termes d'un langage documentaire est normalisé et codifié par rapport à des significations multiples, à des ambiguïtés…

La fonction première d'un langage documentaire est donc une fonction de mise au point qui permet l'exactitude. [4-Maniez]

3 - Fonction d'organisation des connaissances : les classifications et le thésaurus

Le sens de chacun des termes d'un langage documentaire est normalisé et codifié par rapport à un contexte, à un domaine de connaissances…
Un sujet est inséparable de son contexte.
Un langage documentaire permet, par les relations sémantiques qu'il tisse entre les concepts, de naviguer dans un domaine de connaissances.

A travers cette organisation du savoir, la recherche d'informations peut s'effectuer par navigation ou par combinaison entre concept proche, entre l'expression stricte d'un sujet et des expressions voisines ou plus larges. L'utilisateur possède ainsi un outil qui lui permet de mieux réfléchir à sa question, de trouver d'autres angles ou d'autres idées, d'élargir ou de recentrer le cadre de son investigation.

Les langages documentaires ont donc pour deuxième fonction de permettre à la fois l'extension et la spécification d'une recherche. Lorsqu'elle est gérée automatiquement cette fonction s'appelle l'autopostage.
Les langages documentaires sont fondés sur la représentation structurée d'un ou de plusieurs domaines de la connaissance. Ils offrent un accès à la connaissance d'un domaine.

Les classifications, par exemple, obligent à mettre un concept dans un cadre préétabli, c'est-à-dire à poser des repères par rapport à des catégories ou des savoirs généraux. Elles permettent de transposer mentalement une organisation systématique des savoirs rencontrés physiquement, dans l'organisation spatiale des ouvrages dans une bibliothèque ou un centre de documentation.

Les langages de type thésaurus offrent un tableau vivant des concepts utilisés pour décrire un domaine. Comprendre la nature des relations unissant ces concepts entre eux, c'est s'approprier un domaine de connaissances. Cette connaissance permet à l'indexeur d'affiner son analyse de contenu, de mieux percevoir les non dits et les connotations d'un domaine.

Le langage documentaire est donc un outil d'aide à la connaissance d'un domaine et donc le gage d'une analyse de contenu de meilleure qualité pour l'indexeur, et d'une recherche d'informations plus fine, plus adaptée pour l'utilisateur.

4 - Fonction de structuration des documents : typologie et plan de classement

Les langages documentaires permettent également de classer, d'ordonner et de structurer les documents et l'information.
Cette fonction du langage documentaire et de l'indexation ne se situe plus seulement au niveau du sujet et du sens du document mais au niveau de sa forme, de sa structure, du type d'informations qu'il contient.
Dans nombre de recherches, on ne désire pas seulement un sujet précis mais un type d'informations particulier : exemple des chiffres ou un bilan sur le commerce extérieur de la France pour l'année en cours.

Une bonne indexation suppose la structuration préalable du document en plusieurs rubriques (appelées aussi zones ou champs). Chaque rubrique correspond à des catégories d'informations clairement identifiées à l'intérieur du document. Les rubriques de type, date, auteur, titre, support, sont des caractéristiques objectives du document.
Viennent ensuite les caractéristiques subjectives comme le champ sujet qui pourra être subdivisé en thème principal et secondaire afin de mieux structurer l'analyse de contenu.
Cette opération de structuration est en fait une description. Les caractéristiques d'un document sont regroupées dans des notices bibliographiques ou catalographiques.

Selon le domaine où l'on exerce et la nature des documents, d'autres champs, d'autres typologies pourront être définis.
Pour des articles de presse par exemple, on pourra créer des typologies en fonction des pays ou des zones géographiques, des personnes morales ou physiques, de la rubrique du journal (économie, sport, société…), du genre de l'article (entretien, portrait, déclarations, synthèse, enquête, point de vue, etc.) ou du type d'informations (chronologie, statistiques, historique, sondage, texte de loi, etc.).

Chacun des champs sera lié à un index contenant des termes normalisés, de la simple liste alphabétique de mots-clés au langage documentaire le plus complexe, sur lequel s'appuiera l'indexation.

C’est à partir de cette structuration que l’on établit le plan de classement et donc le rangement physique du document.

Celui-ci peut se baser sur une typologie auteur, un ordre alphabétique ou plus souvent sur l’indexation matière d’un thésaurus ou la thématique sujet d’une classification.
Le plan de classement structure les documents par type d’informations, par type de contenu. Le plan de classement donne l’organisation générale d’un fonds documentaire ; il ordonne au préalable l’information ; il permet de localiser l’information ; il est indispensable pour retrouver physiquement un document.

B - Typologie des outils : indexation automatique et traitement linguistique

Le développement des outils informatiques est d'abord venu aider l'indexation humaine et l'utilisation des langages documentaires.
La plupart des logiciels documentaires permettent une gestion informatisée des thésaurus : contrôle des relations entre descripteurs, gestion des candidats descripteurs, etc.
Les logiciels de base de données facilitent la gestion des données textuelles et l'accès à une information structurée : recherche sur une notice avec différents champs auteurs, date, mots-clés, résumés, etc.
L'indexation humaine peut être « assistée par ordinateur » : le logiciel analyse le contenu du texte, le compare à un langage documentaire (thésaurus ou simple liste de mots-clés), fait une proposition d'indexation qui est validée ou modifiée par l'indexeur.
Mais nous sommes là encore dans le contexte d'une analyse préalable des documents, d'une indexation humaine faite le plus souvent à partir d'un langage documentaire.

C'est un autre développement informatique, un autre type d'outils qui nous intéresse ici. Appelé moteur de recherche, moteur d'indexation et de recherche ou infologiciel, ces outils proposent une autre manière d'accéder aux documents et à l'information basée sur l'indexation automatique.

« Indexer automatiquement repose sur deux postulats : les mots du texte sont des indices formels de son contenu et l'exploitation intelligente de ces indices formels constitue une solution alternative à l'indexation humaine… »
[2-Maniez, p°161]

Ce mode d'accès à l'information est apparu lorsque les capacités de calcul et de mémoire des outils informatiques ont permis de stocker et d'analyser des quantités suffisantes de documents.
Ces outils sont d'autant plus présents aujourd'hui que les documents numériques se généralisent, que la masse d'information augmente chaque jour et que cette information est à 80% une information textuelle non structurée.

Dans cette perspective, la recherche d'informations est effectuée sur le contenu du texte, sur le texte intégral. Les techniques d'indexation automatique sont nombreuses et en développement constant : linguistique, sémantique, statistique, etc.

1 - La recherche en texte intégral et l'indexation automatique

A l'expression couramment utilisée recherche plein-texte, qui est une mauvaise traduction de « full-text » en anglais, je préfèrerais le terme de recherche en texte intégral ou sur le texte intégral, c'est-à-dire recherche sur le contenu textuel du document.
L'indexation automatique, pris au sens de constitution d'un index informatique, est un préalable indispensable à la recherche.

L'indexation automatique de base est la constitution d'un index de tous les mots du texte, avec l'adresse de chaque mot dans le document. Les mots-vides ne sont pas éliminés.
L'indexation de premier niveau est fondée sur l'élimination des mots-vides (articles, prépositions, mots grammaticaux…) à partir d'un dictionnaire de termes (appelé « stop list ») et sur la constitution d'un index des termes non éliminés, considérés comme des chaînes de caractères. [14-Sébillot, p°157]

A la recherche, il suffira de taper l'un de ces termes ou une combinaison des termes pour retrouver le document.
Les mots de la requête sont alors simplement comparés aux fichiers d'index des documents.

Cette méthode d'indexation très fruste a de nombreux inconvénients qui limitent très sérieusement la qualité de la recherche d'informations :

- l'absence de prise en compte de l'ordre des mots : gestion politique et politique de gestion seront équivalents
- l'absence de prise en compte de la fréquence et de la position des termes
- l'apparition d'un mot sous ces différentes formes : féminin, pluriel, conjugaison…
- l'analyse porte seulement sur des mots isolés (des unitermes), et délaisse les expressions ou mots composés souvent porteurs de sens : « pomme de terre » sera analysé en deux mots « pomme » et « terre »
- les ambiguïtés du langage naturel ne sont pas traitées : polysémie, synonymie, etc.
- l'obligation d'une stricte équivalence entre termes de recherche et termes indexés pose le problème des fautes de frappes

De nombreux moteurs de recherche dans les bases de données ou sur Internet n'effectuent encore que ce type d'analyse.
La recherche en langage naturel est alors extrêmement aléatoire. L'utilisation d'un langage d'interrogation est indispensable : opérateurs booléens (et ; ou ; sauf) ; troncatures (prise en compte des formes dérivées d'un terme) ; opérateurs d'adjacence (proximité entre deux mots de la recherche), utilisation des guillemets (expressions exactes), etc.

2 - Les traitements linguistiques : analyse morpho-lexicale et syntaxique

Le développement des outils de traitement automatique du langage naturel (TALN) a permis une vraie prise en compte des problèmes d'ambiguïté du langage et une amélioration de la qualité de l'indexation automatique.

La linguistique n'est plus seulement le propre des langages documentaires. On peut distinguer plusieurs niveaux d'analyse linguistique du texte intégral, plusieurs niveaux dans les techniques de TALN :

- le niveau morpho-lexical qui suppose une reconnaissance du mot selon sa forme (réduction du mot à sa forme canonique : lemmatisation)
- le niveau syntaxique qui utilise le filtre de la grammaire
- le niveau sémantique qui tente une reconnaissance des concepts

L'analyse morpho-lexicale ramène chacun des mots d'un texte à une forme normalisée (singulier et infinitif). A cette forme canonique sont associées les formes fléchies du mot (pluriel, féminin, conjugaison) et ses formes dérivées (dangerosité et dangereux liés à danger).
Il existe trois types de formes de base avec les variations liées à cette forme :

- le radical : mang(e) pour manger, mangeoire, mangeables…
- la racine : nation pour nations et nationalité
- le lemme : infinitif des verbes, masculin singulier

Ce processus, appelé lemmatisation, diminue fortement le nombre de mots analysés en éliminant toutes les flexions et les dérivations.

« La recherche peut alors s'effectuer sur une famille de mots, dérivés d'un même radical et présentant de ce fait des parentés de sens : informer, information, informatique, informatisation. » [20-Collas et Chartron, p°11]

L'analyse morpho-lexicale identifie également les groupes nominaux, les expressions idiomatiques et les mots composés : le terme « pomme de terre » sera indexé comme une entité unique de sens. Elle permet le repérage des expressions disjointes : par exemple reconnaître l'expression Agence de presse dans l'expression Agence française de presse.

L'analyse syntaxique d'un texte va permettre de lever des ambiguïtés de sens par une analyse grammaticale.
Les mots seront reconnus et analysés en fonction de leurs catégories grammaticales : nom, verbe, adjectif, etc.
L'homonymie du langage naturel diminue alors fortement puisque l'on perçoit les différences de sens entre, il lit le livre et il livre le lit, entre le « or » conjonction de coordination et le « or » nom commun.

L'approche linguistique n'a été possible que lorsque la puissance de calcul et la capacité de mémoire ont été suffisantes pour la constitution et l'utilisation des énormes ressources que sont les dictionnaires électroniques.

Un dictionnaire électronique est une liste de termes ou expressions avec leurs synonymes et leur appartenance grammaticale. Les termes possibles d'une langue et les expressions idiomatiques ; pour chacune des entrées est donnée la forme grammaticale, la forme canonique, les formes fléchies ou dérivationnelles. [16-Leloup, p°72]

Au-delà de l'analyse linguistique, certains moteurs offrent des possibilités de corrections orthographiques des termes de la requête. Dérivée des techniques de reconnaissance de caractères, la phonétisation permet une recherche sur des mots ressemblants à x caractères près.
Cette technique est particulièrement utile pour des sujets à variations orthographiques comme les noms propres.

Certains moteurs de recherche sont aujourd'hui capables de mettre en œuvre un traitement statistique d'une partie des problèmes linguistiques : lemmatisation et reconnaissance des groupes nominaux par des calculs statistiques à partir d'un corpus de textes, sans avoir recours aux dictionnaires.

3 - Les traitements sémantiques : la sémantique conceptuelle

Les traitements linguistiques ne lèvent qu'une partie des ambiguïtés de sens. L'analyse sémantique va s'intéresser au regroupement de termes synonymes, aux familles de termes, aux réseaux de relations sémantiques entre les termes et dans un texte.

On distingue généralement deux niveaux d'analyse sémantique.
Le premier niveau d'analyse va tenter de prendre en compte les relations sémantiques qui existent entre les termes : la synonymie ; l'hyperonymie (bateau est le père de bateau à moteur) ; la méronymie (voilier appartient à la catégorie bateau) et le renvoi à un autre concept (bateau et construction navale).

A ce niveau d'analyse sémantique, la polysémie peut également être traitée. L'intégration de dictionnaires spécialisés sur un domaine permet de « comprendre » la signification de la fraise non pas comme fruit mais comme outil.

Une analyse du contexte des mots, dans la phrase et dans le texte, rend possible la différenciation entre l'avocat fruit lorsqu'il est proche de termes comme récolte ou agriculture et l'avocat métier lorsqu'il est associé à justice ou tribunal.

Peu de logiciels peuvent aujourd'hui prétendre à ce genre de traitement automatique qui vise à la compréhension d'un mot dans son contexte et à l'énonciation de ses relations sémantiques en tant que concept.
C'est pourtant bien là que se situe le véritable enjeu des traitements automatiques du texte intégral.

Les ressources linguistiques, qu'il est nécessaire d'intégrer pour effectuer ces traitements, sont très importantes : dictionnaire de synonymes ; dictionnaire terminologique (sens en fonction du contexte) ; dictionnaire spécialisé ; dictionnaire de noms propres ; langage documentaire (thésaurus) ; graphes de relations sémantiques entre les concepts ; etc.

Les techniques utilisées marient la statistique et la linguistique dans des algorithmes complexes.

Certains outils se basent sur des dictionnaires, des thésaurus ; d'autres mettent en œuvre des réseaux sémantiques complexes ; d'autres encore fonctionnent par apprentissage à partir d'un corpus de texte représentant un domaine de connaissance.

Le second niveau d'analyse sémantique serait celui de la pragmatique et de l'implicite. *(Voir annexe 1, p°I)*
La véritable compréhension d'un texte va au-delà du sens des mots qui le composent, même si ces mots ne sont pas ambigus.

Déchiffrer les non-dits et les connotations d'un discours nécessite une connaissance du monde, de la société, des savoirs abordés, des situations décrites, ainsi que des facultés d'abstraction et de modélisation propre à l'être humain.

« Il ne faut donc pas s'attendre à ce qu'un programme de traitement du langage naturel puisse extraire d'un texte des concepts qui n'auront pas été prévus par les concepteurs du système, ou qu'il puisse construire des modèles nouveaux de représentation de l'information. » [19-Lefevre, p°44]

4 - Pertinence et classification : de la statistique à la sémantique

Pour qu'une recherche soit vraiment efficace, il est indispensable de sélectionner les documents-réponses en fonction de leur valeur, de leur importance, de leur pertinence par rapport à la question posée.

A tous les niveaux d'indexation automatique que nous venons de décrire, s'ajoute donc des traitements statistiques plus ou moins évolués qui permettent de calculer le poids d'un mot (pondération), c'est-à-dire l'importance d'un mot dans un document déterminé.

Lorsque les mots significatifs d'un texte sont relevés, on calcule statistiquement leur fréquence d'apparition ou occurrence selon un indice moyen. Si l'indice de fréquence d'un mot est supérieur à son indice moyen, on en déduit qu'il doit s'agir d'un mot-clé pertinent, décrivant bien le sujet du texte.

Le premier traitement statistique est alors un calcul de fréquence des mots extrait de la question dans les documents trouvés. Le lot de résultat est trié par pertinence en fonction du nombre de document et du nombre d'occurrence des mots de la requête.

Mais il existe bien d'autres critères de pertinence et de pondération :

- les termes rares : un mot porte d'autant plus d'information qu'il est fréquent dans un petit nombre de texte et rare dans le reste du corpus
- la proximité et l'ordre des termes de la requête dans les documents : plus les mots cherchés sont proches plus ils sont signifiants et plus le document est pertinent
- la position du mot dans le texte : les mots situés dans le titre étant par exemple considérés comme plus important

A tous ces niveaux les algorithmes statistiques peuvent prendre en compte les traitements linguistiques effectués au préalable. Par exemple, pondérer les synonymes d'un mot de la requête : pour voiture, automobile aura 80% de pondération et « formule 1 » 50% de pondération. [16-Leloup, p°73]

Certains moteurs de recherche proposent, en plus du classement par pertinence, une présentation des résultats par classification automatique. Il existe deux méthodes de classification automatique.

La première regroupe les documents dans des classes préétablies, selon des catégories prédéfinies, par un plan de classement par exemple.
La seconde effectue un regroupement des documents en classes à partir de concepts non déterminés au préalable. Ces concepts sont extraits des documents par une analyse de texte basée sur une démarche statistique et linguistique.
Les résultats de recherche sont classés en fonction des termes les plus pertinents présents dans les documents, en fonction des expressions les plus fréquemment associées aux mots qui composent la requête.

En réponse à une requête effectuée à partir du mot clé « vache folle », les documents sont classés dans les catégories : « crise de la vache folle », « maladie à prions », « encéphalopathie spongiforme bovine » ou « maladie de Creutzfeld-Jacob ».

Cette classification automatique est une aide à la recherche : grâce à l'hypertexte, l'utilisateur peut naviguer entre ces différents concepts et ajuster son interrogation.

Enfin, la classification des résultats peut être complétée par un traitement visuel effectué par des logiciels de cartographie. Les groupes de textes sont alors matérialisés par un cercle plus ou moins grand selon leur pertinence, plus ou moins proche de la requête, avec plus ou moins de liens avec les autres groupes de textes.
Ces cartes de connaissances représentent visuellement le contenu informationnel d'un corpus de résultats.

II - PROBLÉMATIQUE ET MÉTHODOLOGIE

A - Langage documentaire et nouvelles technologies : la problématique de l'accès à l'information

Nous avons vu les différentes techniques mises en jeu par les langages documentaires et les technologies informatiques pour indexer, analyser, décrire le contenu d'un document et traiter les ambiguïtés du langage naturel.
L'ensemble de ces processus vise à améliorer la recherche, à faciliter l'accès à l'information. C'est donc à travers les problématiques de la recherche et de l'accès à l'information que l'on pourra évaluer les rôles et fonctions des langages documentaires et des technologies informatiques.

On a longtemps opposé l'indexation automatique et les langages documentaires comme étant des modes d'accès à l'information concurrents. Et en effet, les modes d'accès proposés sont différents, les types d'informations et les stratégies de recherche misent en place ont chacun leurs avantages et leurs inconvénients.

Pourtant, dans les systèmes de recherche d'informations aujourd'hui, il n'y a pas de dichotomie aussi nette entre indexation humaine et utilisation d'un langage documentaire d'une part et indexation automatique et recherche en langage naturel sur le texte intégral de l'autre.
[14-Sébillot] [19-Lefèvre]

La plupart des systèmes en place allient les différentes techniques existantes en fonction des contraintes du contexte : type de documents, domaine couvert par les informations, besoins des utilisateurs, etc.
Pour évaluer et choisir le meilleur mode de recherche, il faut tenir compte de l'ensemble des problèmes posés par l'accès à l'information et adapter les solutions au contexte de recherche et aux besoins des utilisateurs.
[27-Coret, Menon, Schibler et Terrasse]

Nous verrons d'abord les questions et problèmes posés par ces deux types d'accès à l'information. Puis nous définirons les contraintes spécifiques du domaine de la presse et enfin nous tenterons de poser la problématique précise qu'induit ce contexte.

1 - Retrouver un document ou une information : Problématique de la recherche d'informations

L'utilisation de la médiation d'un langage documentaire correspond à un certain type de recherche, à certaines contraintes dans des contextes particuliers. Il en est de même pour l'indexation automatique et les outils de TALN.

Dans un fonds papier, le langage documentaire est incontournable pour organiser, classer et indexer les documents. La mise à disposition de l'information est effectuée à travers le langage choisi par une traduction des questions des utilisateurs. Mais les problèmes de la masse de documents à traiter, de la place prise par le fonds documentaire deviennent vite difficiles à gérer.

Dans un fonds numérisé, il est aujourd'hui possible de stocker et de traiter des quantités importantes de documents. Les modes d'accès à l'information varieront en fonction du type d'indexation choisi.

Opter pour l'indexation manuelle implique que l'on attache du prix à une représentation précise du document. L'indexation humaine est une indexation conceptuelle en profondeur, mais elle n'est pas sans défauts.
En effet l'indexation par concept ou par sujet ne peut être uniforme. Elle varie en fonction de l'indexeur et de sa perception du document, en fonction du temps passé à l'indexation et de biens d'autres critères.
Cette indexation humaine est certes de qualité mais elle n'est pas homogène. De plus, le processus intellectuel de l'indexation est long et donc très coûteux.

Opter pour l'indexation automatique implique que l'on reporte sur le module d'interrogation l'ajustement des documents aux questions posées. L'indexation automatique est constante, mais sa qualité dépend des traitements statistiques et linguistiques mis en œuvre.

« L'indexation automatique présente l'avantage d'une régularité de processus, ce qui constitue une qualité du système, mais qui est différent de la justesse de l'indexation » [11-Le Loarer, p°159]

Si le gain de temps est important, le coût d'un logiciel de TALN n'est pas négligeable.

Les modes d'accès à l'information et les stratégies de recherche varient en fonction de ces deux options.

Lorsque l'indexation s'effectue à partir d'un langage documentaire, se sont les informations secondaires, les métadonnées d'un document qui permettent l'accès au document et à l'information.
La recherche s'effectue sur des champs prédéterminés.
Deux stratégies de recherche sont possibles : recherche par mots-clés, opérateurs booléens et croisement des requêtes ; recherche navigationnelle dans l'arborescence d'un plan de classement et affinage de la requête.

La recherche par mots-clés est une recherche exacte, mais pas exhaustive, qui dépend de la qualité de l'indexation manuelle. La maîtrise du langage d'interrogation est un art difficile réservé aux spécialistes que sont les documentalistes. La mise à disposition de l'information nécessite bien souvent la médiation d'une documentaliste, la traduction de la requête.

L'indexation automatique donne accès à des informations primaires. La recherche s'effectue sur le texte intégral, et l'interrogation est formulée en langage naturel. Deux stratégies de recherche sont possibles : historique des recherches et croisement ; recherche par hypertexte et navigation dans une classification.
La recherche sur le contenu textuel des documents est une recherche de vraisemblance entre la question posée et les documents trouvés qui dépend du type d'indexation effectuée sur les documents et sur les requêtes. [22-Role]
L'utilisation du langage naturel facilite l'accès direct à l'information pour l'utilisateur final. La mise à disposition de l'information ne nécessite pas a priori de médiation.

Si les caractéristiques du fonds documentaire induisent des choix techniques et des types d'organisation en fonction du support et de la quantité de documents, il faut également prendre en compte les moyens humains et financiers, le coût en temps et en argent des systèmes de recherche d'informations choisis.

Mais le choix des différents niveaux d'accès à l'information ne peut s'effectuer qu'après une analyse du contexte. Ce sont les besoins des utilisateurs, le type de questions et d'informations visées, les domaines couverts qui conditionnent le type d'organisation, d'indexation et de médiation documentaire. [10-Fidel]

2 - Le contexte particulier de la presse et ses contraintes

Je m'attacherais ici à définir les caractéristiques particulières de l'information de presse et les contraintes spécifiques liées à l'analyse de contenu.
Un fonds documentaire de presse, qu'il soit papier ou informatique, rassemble principalement des articles en provenance des journaux quotidiens, hebdomadaires ou mensuels, nationaux ou étrangers.

L'information presse est structurée : un article possède un titre, un chapeau, un sous titre, des paragraphes. Cette structure a un sens, elle aide à la compréhension. Il faut donc la préserver lors de l'indexation, qu'elle soit humaine ou informatique.

L'information de presse est redondante dans le temps et dans l'espace.
« Chaque jour qui passe apporte de nouvelles informations sur des événements en cours, ces informations doivent être resituées dans un contexte [...] et différents journaux, édités le même jour couvrent ces évènements. »
[28-Cotte, p°43]

L'information de presse est vite obsolète. On estime qu'en deux ou trois ans, 80% des informations contenues dans une documentation de presse sont périmées.
Le choix de l'information, la sélection des articles et la mise à jour du fonds sont donc des opérations fondamentales.

Mais c'est par rapport à l'analyse de contenu que l'information de presse pose le plus de problème.

En effet le discours de presse est sémantiquement flou. [28-Cotte p°42]
Le même sujet d'actualité sera traité et formulé différemment selon le style narratif du journaliste, la vision et les positions du journal. Le vocabulaire de Libération n'est pas celui du Figaro. Dans le corps d'un article, différents termes seront employés pour désigner le même objet afin d'éviter les répétitions : chef de l'état ou président de la république, Proche-Orient ou Moyen-Orient. Les noms propres, avec leurs écritures multiples, sont très présents dans l'actualité. Le vocabulaire évolue en permanence au gré de l'actualité, et d'antimondialiste on passe à altermondialiste pour désigner le même sujet.

Le discours de presse est donc profondément ambigu, changeant, multiple. La normalisation, la levée des ambiguïtés du langage, la représentation du savoir sont des conditions sine qua non du traitement de l'information presse.

Enfin, les besoins des journalistes doivent être pris en compte pour choisir la meilleure indexation et le mode de recherche le plus adapté. On distingue quatre types de demandes et de recherche d'informations.

- Besoin précis en information factuelle pour vérification : une date, un lieu, un personnage, un événement
- Besoin précis d'un document ou d'un type de document : un article à retrouver, un type de document (portrait)
- Besoin d'un sujet ou d'un thème : état de l'art sur une question
- Besoin d'une idée ou d'un angle : feuilletage

Les besoins des journalistes sont divers et complexes : un document précis, une information pointue, mais aussi des synthèses sur des sujets plus larges, des articles dont l'orientation est originale, et même des idées.
Or, l'information presse a ceci de particulier qu'elle couvre en fait la totalité de l'activité humaine : économie, politique, culture, international, scientifique, etc.
Pour satisfaire les besoins des journalistes, l'indexation devra être à la fois large, pour préserver une organisation généraliste du fonds documentaire, et précise, pour permettre une analyse spécifique des documents et une hiérarchisation de l'information. [28-Cotte p°34]

3 - La question posée

Il s'agit de définir le rôle, la place et les fonctions du langage documentaire aujourd'hui, sa forme et ses spécificités face aux performances des outils informatiques et linguistiques.

Pour tenter d'évaluer ces rôles et fonctions dans la recherche d'informations et l'accès à l'information, nous partirons du contexte de la presse et des besoins des journalistes.
Nous avons vu que les particularités de l'information de presse imposent une normalisation, une désambiguïsation du langage naturel et une forte organisation des connaissances.

L'évaluation des résultats d'une recherche s'effectue, quelle que soit la technique utilisée, en fonction de deux critères : le bruit et le silence.

La mesure du bruit ou taux de précision : Le nombre de documents pertinents retrouvés par rapport au nombre total de documents retrouvés.
Plus le taux de précision est faible, plus on est confronté au phénomène du bruit.
La mesure du silence ou taux de rappel : Le nombre de documents pertinents retrouvés par rapport au nombre total de documents pertinents.
Plus le taux de rappel est faible, plus on est confronté au phénomène du silence.

« Tout effort pour augmenter le taux de rappel se paie par le risque de diminuer parallèlement le taux de précision à l'indexation comme à l'interrogation. »
[2-Maniez, p°184-185]

Une indexation avec de nombreux descripteurs est source de bruit, de même que l'indexation automatique. Faire une indexation plus légère et appliquer les procédés de normalisation et de désambiguïsation du langage naturel permet de diminuer ce bruit.
A l'interrogation, l'accumulation de mots dans une requête est source de silence.

C'est de l'indexation que dépendent principalement ces phénomènes de bruit et de silence. Et c'est en fonction des besoins des utilisateurs, du type de questions posées que l'on définira les traitements et l'indexation qui s'imposent.

Nous pouvons donc poser une typologie de l'indexation à partir de laquelle nous décrirons les possibilités des langages documentaires et des nouvelles technologies :

- Besoin précis, recoupement et vérification d'une date, d'un fait : une indexation descriptive et factuelle
- Besoin précis d'un article, d'un type d'article : une indexation fonctionnelle pour une structuration des articles et une typologie de l'information
- Besoin d'une synthèse sur un thème, d'un état de l'art sur une question : une indexation thématique et analytique pour une définition du sujet et des concepts

- Besoin de feuilletage et boite à idées : la fonction de navigation par le langage ou par l'hypertexte et la classification.

C'est par rapport à cette typologie de l'indexation que nous comparerons les possibilités offertes par les langages documentaires et les outils informatiques. Nous tenterons d'évaluer les qualités et défauts de la recherche d'informations et des stratégies de recherche, la pertinence et la validité des résultats.

B - Observation chez Bayard

Lors de mon stage au centre de documentation de Bayard, j'ai pu observer sur le terrain l'ensemble du fonctionnement d'un centre de presse.
Après une présentation rapide, je décrirais plus précisément les points qui me permettront d'illustrer mon étude et de mieux comprendre les particularités d'un centre de documentation dans la presse aujourd'hui.
[23-Briand] [29-Eyraud]

Le centre de documentation emploie 9 personnes qui travaillent uniquement pour les salariés du groupe, principalement des journalistes.
Le fonds documentaire est un fonds papier composé de 100 000 dossiers documentaires thématiques (consultation sur place uniquement), une collection de 200 magazines, 20 000 ouvrages d'actualité et de références et les collections des revues du groupe (La Croix sur microfilm depuis 1883).
Les documentalistes sélectionnent et indexent les articles de 10 quotidiens et de plusieurs dizaines d'hebdomadaires et de mensuels pour alimenter les dossiers documentaires.

Les utilisateurs sont accueillis sur place mais peuvent également poser des questions par téléphone. Les 40/45 questions journalières en moyenne regroupent les consultations, les questions précises et les dossiers sur mesure effectués à la demande. Pour cela, les documentalistes utilisent l'ensemble des ressources du centre et interrogent des banques de données de presse en externe (Européenne de données, Europresse, Factiva, etc.)
Plusieurs revues de presse spécialisées, dont une quotidienne, et un agenda prévisionnel complètent la gamme de service du centre.
Sous format informatique, le journal La Croix est indexé chaque jour avant exportation vers l'Européenne de données et Europresse.

1 - Le rôle du langage chez Bayard et son évolution

Le langage de Bayard est très particulier. C'était au départ, il y a plus de dix ans, une liste alphabétique de descripteurs servant pour l'indexation matière et comme plan de classement des dossiers.

De nombreux centres de presse ont choisi cette organisation où le langage d'indexation correspond au rangement physique des documents, à la structuration des dossiers thématiques.

Au fil du temps, ce langage s'est peu à peu développé, transformé, adapté aux circonstances et à l'actualité.
Aujourd'hui, sans être tout à fait un thésaurus, il en a certaines des caractéristiques : liste de renvois, notes d'application, relations d'associations et relations hiérarchiques.
Il s'organise aujourd'hui autour de six grands pôles : biographies, entreprise, matière (c'est à dire sujet France), géographie (pays, zone géographique, organisations internationales, etc.), géographie France (villes, départements, régions, fleuves, etc.) et médias (généralités, presse, audiovisuel, radio, TV, Bayard presse, La Croix). Chacun de ces pôles correspond à une zone de classement, chaque descripteur correspond à un dossier.
(Voir annexe 2 p°II, III et IV)

Les dossiers des pays comportent par exemple une vingtaine de sous thèmes : économie, agriculture, politique étrangère, culture, etc.
Dans la liste matière, le sujet des retraites est subdivisé en plusieurs dossiers :
RETRAITE/GENERALITES, RETRAITE/LEGISLATION-REFORME, RETRAITE/CALCUL, RETRAITE/TRAVAIL, RETRAITE/PAYS, etc.
Le langage s'est peu à peu affiné et hiérarchisé pour pouvoir répondre aux questions d'un public de plus en plus nombreux ayant des centres d'intérêts hétérogènes et aux exigences de l'actualité.

La gestion et la mise à jour du langage documentaire sont fondamentales. De nouveaux descripteurs sont intégrés, la liste de renvois est enrichie constamment, des conseils d'indexation précisent la signification des descripteurs et le contenu des dossiers.
Les règles d'écritures sont précises notamment pour les noms propres des dossiers biographiques.

Le langage documentaire organise les domaines de connaissance présents dans l'actualité ; il normalise et définit les termes employés pour éviter les ambiguïtés.

Il permet la traduction des requêtes des journalistes et l'accès direct aux dossiers qui sont la première source d'information des documentalistes.

Il permet également de mettre en valeur et de préserver les centres d'intérêts du groupe Bayard à savoir les jeunes, les seniors et la religion. Ces trois thèmes sont donc privilégiés lors de la collecte d'information et grâce à une indexation plus précise.

La sélection des articles s'effectue également en fonction des sources. Tous les articles de La Croix sont dépouillés et indexés et Le Monde est conservé presque intégralement.
On indexe un article avec deux ou trois mots-clés au maximum afin de préserver la lisibilité et la spécificité des dossiers et donc d'éviter trop de bruit à la recherche.

2 - Du papier à l'informatique

Les 100 000 dossiers thématiques sous format papier représentent pour Bayard une véritable richesse. Ils permettent de répondre à près de 70% des questions des utilisateurs.
Mais les problèmes de gestion d'un fonds papier sont de plus en plus prégnants : la place prise par le fonds documentaire, le temps et le coût de l'indexation, la gestion lourde des dossiers (bourrage), etc.
Chez Bayard, comme dans la majorité des autres centres de documentation presse, se pose donc le problème de l'informatisation.
J'ai participé durant mon stage à cette réflexion sur l'informatisation et aux diverses solutions possibles notamment en matière d'intégration du langage.

Bien sûr l'informatique est depuis longtemps présente dans les centres de presse. [23-Briand]

La plupart des journaux ont commencé par informatiser leur propre titre dans une base de données, qui leur permettait d'intégrer leur langage, de préserver un lien avec le dépouillement papier, d'indexer et de structurer les articles sélectionnés et de proposer des stratégies de recherche sur les métadonnées.
De plus, de nombreuses bases de données de presse sont interrogées en externe. Ces bases fonctionnent en combinant une recherche en texte intégral, une interrogation par date ou source et des équations booléennes sur le titre et le chapeau.
J'ai pu observer et tester moi-même les résultats d'une recherche avec de tels outils.

En dehors de l'archivage de leur propre titre, les centres de presse continuaient et continuent encore le dépouillement papier des autres journaux.

Mais aujourd'hui, la situation a changé. La puissance des outils permet de traiter des informations textuelles en grande quantité. Les traitements linguistiques et statistiques peuvent prendre en charge les ambiguïtés du discours de presse. Mais surtout un accord de coopération documentaire s'est mis en place entre les centres de documentation des journaux de la presse quotidienne nationale. Cet accord a été signé entre Libération, Le Monde, Le Figaro, La Croix, Les Échos et La Tribune et est effectif depuis un an. Le même type d'accord existe entre l'Express, Le Nouvel Observateur, Le Point et Télérama.

La coopération prévoit l'échange gratuit des fichiers électroniques des journaux, au jour le jour. Les fichiers en format ASCII[1] balisé proviennent directement du système éditorial de chacun des journaux concernés. 6 à 8 balises sont renseignées afin de transmettre des articles déjà structurés avec la source (nom du journal), la date, le nom de l'auteur, le titre de l'article, sa taille, sa place dans le journal, etc.
Aujourd'hui des problèmes de récupération de données et de compatibilité entre les systèmes persistent.

Dans le cadre de la coopération documentaire, il est donc prévu de passer à un format XML[2] et de déterminer une DTD[3] commune, c'est à dire une description commune et formalisée des documents.
L'utilisation du format XML facilitera à la fois la normalisation et la structuration des documents et permettra une meilleure récupération de ces données.

3 - Les besoins particuliers des journalistes

90% des utilisateurs du centre de documentation sont des journalistes qui appartiennent aux différents journaux et publications du groupe.
Le centre de documentation fait une analyse statistique des questions posées par les journalistes et du type de recherche effectuée pour eux :

- Besoin précis de vérification d'une date, d'un événement et lecture d'un article précis : Question-réponse et référence articles 31%. Cette catégorie correspond en fait à deux types de besoins des journalistes mais nécessite le même type de recherche plutôt courte
- Besoin d'un état de l'art sur une question, le point sur un thème : dossier sur mesure 20 à 23%
- Besoin plus flou, recherche d'une idée ou d'un angle, feuilletage : auto consultation des dossiers 47%.

Les recherches des deux premières catégories sont entièrement prises en charge par les documentalistes. Les journalistes ont rarement le temps même pour retrouver un de leurs articles. La médiation documentaire est donc très importante dans la mise à disposition de l'information.
Dans le cas d'une recherche d'idée, le journaliste préfère souvent feuilleter seul un dossier documentaire. Néanmoins au cours de cette lecture, le dialogue avec les documentalistes est permanent.

Même si avec l'utilisation d'Internet, les journalistes effectuent eux-mêmes une partie des recherches, ils préfèrent en majorité poser leurs questions aux documentalistes.

[1] ASCII : American Standard Code for Information Interchange ou code américain standard pour l'échange d'information

[2] XML : Extensible Markup Language ou langage de balisage extensible

[3] DTD : Définition de types de documents

De plus, les journalistes sont très attachés à la lecture papier des dossiers et des revues de presse.

L'informatisation posera donc le problème de la préservation de la médiation documentaire et de l'adaptation de pratiques de lectures difficiles à transposer.

Difficile aussi de penser une informatisation où l'utilisateur principal serait certes la documentaliste mais où le journaliste pourrait également chercher des informations.
Enfin restent les difficultés à transposer des produits documentaires du papier à l'informatique.

C - Les entretiens menés dans d'autres structures de presse

J'ai complété mes observations chez Bayard par une série de visites dans trois autres centres documentaires de presse : Le Monde, Les Échos et le Nouvel Observateur. Dans ces trois centres de documentation, le fonds documentaire a été, au moins en partie, informatisé. Il était primordial pour moi de pouvoir mener un entretien avec le responsable de la documentation pour comprendre ces choix par rapport à l'informatisation. Mais aussi d'avoir l'occasion d'observer et de tester les bases de données de presse mises en place, les technologies choisies, la place laissée au langage documentaire et les performances de la recherche d'informations.

Ces visites ont été très riches et m'ont permis d'approfondir mon questionnement.
Chacun de ces centres de presse a fait un choix différent en matière d'informatisation et d'intégration du langage. Mais tous ces choix allient recherche sur le texte intégral et indexation humaine. Dans les trois centres que j'ai pu visiter, l'informatisation n'est pas complètement finalisée.
Je ne décrirais dans cette partie que les caractéristiques principales des choix effectués. Les comptes rendus de chacune de mes visites sont placés en annexes.

1 - Le Monde – (Voir annexe 3 p°V à XI)

Depuis le 24 décembre 2002, la nouvelle base de données du Monde est en test à la documentation. La base de données contient plus d'un million de documents (l'intégralité du journal et de ses suppléments depuis 1987).
Les journalistes du Monde ont surtout besoin du Monde comme référence, les autres journaux subissent donc un dépouillement papier très sélectif.
La base de données est un produit sur mesure issu de la coopération entre deux grandes sociétés : Xylème pour son savoir-faire en matière de base de données (format XML) ; Sinéqua pour son moteur de recherche linguistique Intuition et pour son expérience dans la presse.
L'ensemble des documents est indexé automatiquement par le moteur après « le filtrage » des traitements linguistiques : gestion de la synonymie et de l'ambiguïté.

Les dictionnaires intégrés sont la grande force d'Intuition : dictionnaire de la langue française et dictionnaire des synonymes avec une mise à jour tous les six mois en collaboration avec le responsable du centre de documentation.
200 articles par jour entrent dans la base de données. La structuration en XML de la base de données permet de « récupérer automatiquement » le titre, chapeau, auteur, place dans le journal et longueur de l'article.
A chacun des articles est rattaché l'ensemble des images : photographie et infographie (carte, tableau). Ces images sont disponibles en JPG, et le module d'interrogation permet la recherche de ces images.

Une autre image accompagne chacun des articles : un format PDF qui donne la vraie mise en page de l'article dans le journal. Il est très important de préserver cette mise en scène de l'article dans la page.

L'ensemble des articles du journal est marqué à l'entrée dans la base de données. Le responsable du centre de documentation parle de marquage et pas d'indexation, sans doute dans la mesure ou ce marquage est relativement sommaire ; en tout cas sans aucune mesure avec les 30 champs d'indexation de l'ancienne base de données.
Les champs les plus importants sont :

1. Catégories : une liste de 50 catégories qui permettent de faire une typologie de forme des articles (éditorial, brève, etc.) et une typologie de contenu (entretien, portrait, enquête, déclarations, chiffres, chronologie etc.).
2. Titres complémentaires : le titre de l'article est enrichi par les documentalistes. Quelques lignes seulement pour lever l'ambiguïté des titres de presse et permettre une recherche sur la titraille (titre + sous titre + titre enrichi).
3. Genre œuvre : la notion d'œuvre correspond aux critiques. Dans le genre de l'œuvre on trouve cinéma, littérature, théâtre, etc.
4. Titre œuvre : le titre exact de l'œuvre
5. Auteur œuvre : le nom et le prénom de l'auteur

Ces catégories se retrouvent à la recherche et permettent d'affiner le questionnement. La recherche est également possible sur le texte intégral. Le lot de résultat est trié par pertinence (paramétrable à la recherche) et présente une classification dans laquelle on peut naviguer.

2 - Les Échos – (Voir annexe 4 p°XII à XIV)

Il existe deux bases de données séparées gérées par le même outil (Basis+), une pour les archives du journal et l'autre pour la coopération documentaire.
La base de données des archives des Échos comprend les archives du journal depuis 1991. Les fichiers sont en format HTML et sont disponibles pour les journalistes sur l'Intranet.
Le moteur de recherche intégré permet de faire des recherches booléennes classiques sur le titre ou le texte intégral, de croiser des requêtes, de chercher par auteur, date, rubriques du journal, typologie d'articles, etc.
Tous les articles du journal sont conservés ainsi que les infographies qui sont très importantes aux Échos (tableaux statistiques, courbes financières, etc.)

L'indexation de la base est faite manuellement et correspond à 70% au langage documentaire du centre.
De plus en plus il y a un allégement de l'indexation sur la base des échos, avec un souci d'une évolution vers le plus précis. Les infographies sont indexées par un documentaliste.

La base de la PQN[4] est séparée de celle des archives du journal. Elle est gérée par le même logiciel Basis+. C'est une base en format de fichiers HTML, dont l'ergonomie est très simple.
Une sélection est effectuée sur l'ensemble des fichiers de la PQN, mais cette sélection est beaucoup moins importante que celle effectuée pour le dépouillement papier.
Aucune indexation n'est effectuée pour l'instant sur la PQN. Les documentalistes assurent un enrichissement très rapide du titre. Pour les brèves, les 20 premiers mots du texte apparaissent automatiquement dans le cadre du titre, et le documentaliste choisi de garder ou d'enrichir ce choix.

L'objectif, à terme, est de constituer des dossiers virtuels. Une indexation très simple sera effectuée, deux mots-clés au maximum, pour créer des dossiers en ligne qui auront un lien avec les archives des Échos. Ce projet est encore au stade de la réflexion.
Cette base de PQN n'est pas interrogeable par les journalistes. Ce sont les documentalistes qui font des recherches pour eux et qui impriment les articles si besoin. Les recherches sur la base se font uniquement en texte intégral.

La presse quotidienne continue d'être dépouillée en papier dans la mesure où le système de coopération documentaire n'est pas encore complètement au point.

[4] PQN : presse quotidienne nationale

3 - Le Nouvel Observateur – (Voir annexe 5 p°XV à XVII)

Là aussi c'est pour pouvoir exploiter les fichiers de la coopération documentaire, que le service de documentation s'est informatisé.
Le choix s'est porté sur la société Eurocortex et sur un logiciel documentaire de gestion de contenu. Après plusieurs mois de test, la base de données gérée par le logiciel ICM (Intelligent Content Manager) fonctionne depuis mars 2003.

La perspective d'informatisation du service de documentation a nécessité la création d'un thésaurus spécifique. En effet, jusqu'alors l'indexation était effectuée à l'aide d'une « Nomenclature par secteur » qui correspondait au plan de classement physique des documents, c'est-à-dire aux dossiers thématiques. Ce langage très simple est encore aujourd'hui utilisé par l'archiviste pour indexer le Nouvel Observateur.
8 mois de travail ont été nécessaire pour constituer cet outil d'indexation et de recherche en ligne.

Les documentalistes sélectionnent et indexent l'ensemble des documents entrant dans la base de données, en fonction de leurs domaines de spécialisation. La sélection s'effectue sur les hebdomadaires appartenant à la coopération documentaire dont les articles sont récupérés automatiquement en format ASCII balisé. Les champs récupérés automatiquement sont : la source, la date, l'auteur, la page, la taille (en signes), le titre et le chapeau. Comme aux Échos, cette récupération nécessite une vérification et un nettoyage.

Pour les autres journaux, notamment les quotidiens, les documentalistes vont chercher sur les sites en ligne, les articles dont elles ont besoin. Le dépouillement papier est de plus en plus succinct, il concerne les publications quotidiennes et étrangères qui ne peuvent être récupérées en ligne.
A noter que la sélection des articles est très importante ; seuls sont conservés les articles d'une certaine importance, les synthèses, les portraits et interviews.

L'indexation humaine est assez importante. Elle comprend 11 champs : rubrique, personnalité, entreprise, région du monde, pays, région, ville, mot outil, thème (3 champs).
A noter que dans la presse, les mots outils contiennent également une typologie des informations et une typologie des articles : chronologie, statistiques, déclarations, entretien, etc.
Le thème correspond au sujet de l'article. L'indexation est faite avec l'aide du thésaurus en ligne. En général deux ou trois termes sont sélectionnés en essayant d'aller vers le plus précis.

La recherche peut s'effectuer sur les champs de l'indexation ou sur le texte intégral indexé automatiquement. La recherche simplifiée permet aux journalistes de faire une requête sur le titre et chapeau, d'interroger sur l'auteur et la date et de poser une question sur le texte intégral. 30% des journalistes utilisent cette base, principalement pour des recherches simples comme retrouver un article.
La recherche avancée n'est accessible qu'aux documentalistes. Tous les champs de l'indexation sont paramétrables à la recherche.

III-LES LANGAGES DOCUMENTAIRES EN ÉVOLUTION AU CŒUR DES SYSTÈMES D'INFORMATION

Pour rester proche de ce que j'ai pu observer sur le terrain, et pour pouvoir mieux comparer et évaluer les performances des langages documentaires et les possibilités des outils informatiques, nous nous placerons dans un système d'information presse alliant les deux techniques d'indexation et de recherche.

Nous verrons comment, dans une base de données de presse, le langage peut s'intégrer de façons différentes aux systèmes d'indexation automatique du texte intégral.
Soit, on ajoute une indexation humaine dans un certain nombre de champs et on parle alors de surindexation[5]. Cette surindexation peut aller d'un simple choix de quelques termes dans une liste normalisée, jusqu'à l'intégration complète d'un thésaurus.
Soit le langage est intégré en amont, et l'indexation automatique s'appuie directement sur le langage qu'il soit sous la forme d'un dictionnaire métier, d'une liste de renvois, d'un plan de classement, etc.

Selon les systèmes mis en place, il est possible de cumuler, de croiser différents types de recherche en fonction des traitements effectués.
[15-Chaumier et Dejean]

En partant d'une typologie des besoins d'information des journalistes, nous verrons quel est le type d'indexation le plus adapté. Nous tenterons d'analyser les performances de la recherche en matière de bruit et de silence et de pertinence de l'information. Nous pourrons évaluer ce que nous apporte la technique et quelle est la place des langages documentaires, quelles sont leur forme et leur fonction au cœur des systèmes d'information de presse.

[5] Le terme surindexation est employé ici dans le sens de l'ajout d'une indexation humaine à un système d'indexation automatique.

A - Besoins précis, recoupement et vérification d'une date, d'un fait : une indexation descriptive et factuelle moins utile ?

Les recherches concernées sont les recherches d'informations précises non pas sur un sujet général mais sur une personne physique ou morale, un événement particulier, un lieu, un pays.
Exemple de questions : La date de la mort du président Mitterrand, le chiffre d'affaires de l'entreprise Orange, les dernières déclarations de l'Italie à l'ONU sur la guerre en Irak, les attentats de Paris en 1995, etc.
Il s'agit donc de trouver des documents susceptibles de répondre à une question précise. Dans ce cas, la recherche s'effectue principalement sur les mots. L'indexation devra donc être descriptive et factuelle et répondre aux difficultés liées à la normalisation et à l'ambiguïté du langage naturel.
Une question précise ne nécessite pas une recherche approfondie mais seulement quelques documents pertinents.
Des stratégies de recherche simple doivent permettre de répondre rapidement à ce type de question.

1 - Les apports du texte intégral et des traitements linguistiques

Lorsqu'il s'agit de retrouver un renseignement précis de ce type, l'indexation nécessaire se doit donc de décrire et de recenser les faits présents dans un document. Une indexation humaine passera en revue les personnes morales ou physiques, les lieux, etc. en sélectionnant ce qui est central dans l'article.

Mais ces divers éléments sont littéralement déjà présents dans le texte intégral, et une indexation automatique simple permet de les retrouver sans l'apport d'une surindexation. En effet, dans un article sur la mort de Mitterrand, les mots mort et Mitterrand seront présents.

Bien sur, en l'absence de tous traitements statistiques ou linguistiques, le bruit et le silence générés seront considérables. Rechercher sur l'ensemble du contenu du texte, c'est mener une recherche exhaustive. Cela reviendra à avoir dans les résultats des articles sur Danièle Mitterrand, une déclaration de Mitterrand sur la mort de M. X et à ne pas trouver un article sur le décès de l'ancien président de la république. Chacun de nous à pu observer ces travers lors d'une recherche sur Internet.

La maîtrise des équations booléennes permet aux spécialistes de diminuer le bruit, mais ne règle pas tous les problèmes et limite la mise à disposition de l'information.

Pour se passer d'une indexation manuelle sur ce genre de questions, les techniques statistiques et linguistiques doivent être suffisamment efficaces pour diminuer notablement le bruit et le silence.

La mise en place de traitements statistiques, basée par exemple sur la proximité des mots de la recherche, permettra un classement par pertinence. Les articles les plus proches de la question posée apparaîtront en premier, la gestion du bruit sera moins problématique que dans un classement par date par exemple. L'utilisation du langage naturel à l'interrogation rend la recherche plus accessible.

Nous verrons également comment la mise en place d'une classification automatique peut améliorer encore ces fonctions. *(Voir p°43-44)*

L'ajout de traitements linguistiques permet de diminuer les ambiguïtés du langage naturel et par corollaire le bruit et le silence généré par une indexation automatique. [17-Théret]

Pour une question sur les attentats de Paris, l'analyse grammaticale permettra de retrouver les articles où le nom de Paris est présent sans le confondre avec une forme du verbe parier.
Avec un traitement morpho-lexical, le terme « chiffres d'affaires » sera compris comme une expression et cherché en tant que tel.
Dans ces deux cas le bruit diminue fortement.

Un moteur de recherche capable de traiter la synonymie, donnera la possibilité d'étendre la recherche du terme « déclarations » à des équivalents comme proclamations, annonce ou discours.
Grâce à la lemmatisation, le moteur de recherche associe à « déclaration » son dérivé pluriel et les formes du verbe déclarer.
Dans ces deux cas, c'est le silence qui diminue.

A noter que pour que les traitements linguistiques soient véritablement efficaces, la question doit être réellement formulée en langage naturel, sous la forme d'une phrase ou au moins d'une expression.
[24-Dalbin et Saleras, p°317-318]

Selon les paramétrages des moteurs de recherche linguistiques, il peut être important de respecter la casse (majuscule, minuscule) et les accentuations, dans la mesure où ils sont porteurs de sens.

Mais nous allons voir que tous les problèmes ne sont pas réglés par les techniques statistiques et linguistiques appliquées à la recherche en texte intégral.

2 - Synonymie, homonymie, ambiguïté du langage naturel : recherche d'un nom, d'une date, d'un lieu, et recherche rétrospective.

La recherche sur les noms propres, particulièrement importante dans la presse, pose des problèmes spécifiques. [6-Amar, p°268 à273]
Tout d'abord le succès d'une recherche suppose que l'on a utilisé la bonne orthographe et que l'on n'a pas fait de fautes de frappes. Dans les articles de presse, les noms des personnes sont souvent employés sans les prénoms ce qui n'aide pas la recherche.

Les orthographes varient d'un journal à l'autre, d'un journaliste à l'autre. La façon d'écrire Mao Tsé Toung peut se décliner à l'infini. L'orthographe des noms propres étrangers varient énormément comme Al Quaida par exemple.

Une phonétisation peut palier un certain nombre de problème mais sa mise en place est difficile et la tolérance aux fautes de frappes est généralement limitée à deux ou trois caractères. Reste la solution du dictionnaire des noms propres, mais c'est une intégration lourde qui demande une mise à jour régulière par rapport à l'actualité.

L'analyse linguistique ne permet pas de traiter tous les cas d'homonymie.
La recherche des noms des personnes morales notamment, pose d'énormes difficultés. Pour retrouver les articles à propos de l'entreprise Orange, l'analyse grammaticale ne suffit pas. Dans la phrase « l'entreprise orange a annoncé son chiffre d'affaires », rien ne permet de différencier l'adjectif orange et Orange le nom propre. Le logiciel mis en place au Monde permet un paramétrage aux majuscules afin de faire cette différence.

Les sigles sont également complexes à reconnaître pour un moteur de recherche, de même que les dates. Les règles d'écriture varient : ONU avec point, avec espace ou sans, avec développé ou sans, en majuscule, etc.

Une lemmatisation appliquée à des noms propres peut permettre de retrouver Italie et ses déclinaisons italiens et italiennes. Cependant ce type de traitement appliqué aux noms propres génèrera un bruit très important.

Enfin, aucune analyse linguistique ne peut différencier deux articles dont l'un serait une déclaration de Jospin sur Chirac et l'autre une déclaration de Chirac sur Jospin.

Dernier point : l'évolution du discours de presse. La prise en compte des changements de noms, que ce soit pour un pays, une entreprise, un événement ou un mouvement, est indispensable pour effectuer une recherche rétrospective. Pour un journaliste qui s'intéresse au mouvement antimondialiste, la prise en compte de la nouvelle expression altermondialiste est fondamentale.

Or, la gestion de la synonymie nécessite l'intégration d'un dictionnaire.

Cependant, l'intégration d'un dictionnaire de synonymes de la langue courante ne suffira pas à gérer l'évolution du vocabulaire de l'actualité. Ce type de dictionnaire ne permet que la prise en compte de la synonymie « simple ».

Pour être plus efficace, il devra être mis à jour régulièrement, traiter les noms propres, être enrichi avec l'aide des documentalistes qui perçoivent l'évolution du discours. C'est la solution mise en place au journal Le Monde.
L'autre solution consiste à intégrer une liste de renvois ou les liens d'équivalence d'un thésaurus.

Dictionnaire de synonymes ou liste de renvois, c'est donc bien l'utilisation d'une partie d'un langage documentaire qui permettra une véritable recherche rétrospective.

3 - Du bruit au silence : la normalisation du langage documentaire en complément du texte intégral et des traitements linguistiques

Nous avons vu qu'il demeure certaines difficultés liées à la recherche des noms propres principalement pour les personnes morales ou physiques et dans une moindre mesure pour les lieux. C'est donc à ce niveau qu'il sera utile d'intégrer un langage documentaire et une indexation manuelle. Cette intégration permettra une normalisation et donc une pertinence plus importante sur ce genre de recherche.

Normaliser les règles d'écriture des noms propres empêchera, à la recherche, les biais des différences orthographiques.
Si l'on choisit d'ajouter un champ personnes morales ou physiques, seuls les documents réellement importants seront sélectionnés et indexés, le bruit diminuera et la recherche sera plus pertinente.
On peut également choisir d'intégrer une liste des noms propres et des personnes morales les plus importantes en fonction du domaine, et ainsi optimiser les traitements linguistiques.

Ajouter un champ pour les noms de pays peut paraître moins fondamental. Néanmoins, la recherche sur un pays entraîne un bruit considérable. Pour une requête sur Hongkong par exemple, un article comportant la phrase, « de notre envoyé spécial à Hongkong », sera retrouvé, mais aura peu de chance d'être pertinent.

Une surindexation, relativement légère, aura des conséquences très bénéfiques pour la recherche et évitera les problèmes posés par la lemmatisation des noms propres.

De plus, avec une indexation humaine, il sera plus facile de suivre l'évolution du vocabulaire de presse, d'enrichir une liste de renvois et des liens d'équivalence, et ainsi d'optimiser la recherche rétrospective.

Même pour ce type de besoins simples, une indexation manuelle permet d'améliorer la recherche. Reste à savoir si elle est indispensable.
Certes, une part de bruit et de silence persiste avec une indexation automatique, malgré les traitements statistiques et linguistiques. Mais l'intégration d'un dictionnaire de synonymes et de quelques listes normalisées peut suffire sur des questions simples qui ne nécessitent que quelques documents pertinents.

Il s'agit en fait de faire un choix entre, d'un coté, un gain de temps et le problème de la gestion du bruit et du silence, et de l'autre, une indexation et une recherche plus exacte mais un coût en temps et en ressources humaines non négligeables.

De plus, nous verrons qu'une indexation fonctionnelle, plus simple et plus légère, peut permettre de faciliter ce genre de recherche.

B - Besoins précis d'un article, d'un type d'article, d'un type d'information : une indexation fonctionnelle indispensable

Le besoin de retrouver un article précis, un article connu est une requête fréquente de la part des journalistes, et c'est celle qui paraît la plus simple.
Retrouver un type d'articles ou d'informations est une demande qui vient généralement en complément d'une recherche factuelle ou thématique.
Ce type de recherche nécessite la structuration d'un document, son appartenance à un genre et la caractérisation des informations contenues dans un article. Une indexation fonctionnelle permettra de répondre à ce type de besoin.

1 – Recherche d'un document : retrouver un article connu

La recherche précise d'un document implique que l'on puisse retrouver ce document par ces caractéristiques objectives : auteur, date, source. Ce qui signifie pouvoir faire une recherche sur des documents structurés.
Cette structuration peut se faire manuellement par la création et le remplissage de champs prédéterminés, ceux utilisés classiquement dans une notice bibliographique.

Mais, certains langages informatiques permettent dans une base de données, de récupérer automatiquement un document structuré. Dans le cadre de la coopération documentaire, le format ASCII balisé est utilisé pour lier les caractéristiques objectives d'un article au document lui-même.
Cette opération est effectuée au niveau du système éditorial du journal. Pour chaque article on pourra obtenir la source (le nom du journal), sa date, le nom de l'auteur, le titre de l'article, sa longueur, etc.
A noter que ce système paraît bien fonctionner, même si la récupération n'est pas encore parfaite. Aux Échos par exemple, les documentalistes vérifient la validité des informations entrées et procèdent à un nettoyage. Des problèmes de compatibilité entre les différents systèmes persistent.

Si l'indexation humaine n'est plus nécessaire, une vérification demeure indispensable. Néanmoins, le format des articles dans la base de données permet une recherche sur l'auteur, le titre, le support, etc. Cette recherche d'un document précis est accessible aux journalistes et paraît très utilisée.

Reste que cette recherche de base n'est simple qu'en apparence. En effet, le journaliste en feuilletant la presse, repère un papier intéressant mais ne mémorise pas toujours l'ensemble de ses caractéristiques.
Bien souvent la demande est formulée très vaguement : un édito de La Croix, un rebond de Libé la semaine dernière, un papier des pages économiques du Figaro. C'est à partir d'une visualisation du journal et des ces rubriques que l'article est décrit par le journaliste.

Il est donc important d'avoir accès, dans la structuration, à l'appartenance d'un article à une rubrique ou à une zone thématique du journal.
Dans la base de données du journal Le Monde, ces deux typologies sont récupérées automatiquement en extraction XML. On peut donc, à la recherche, choisir de faire une requête sur les articles des pages internationales du monde par exemple.

Sans une structuration de ce type, c'est par l'intégration manuelle d'une première typologie que l'on pourra répondre à ce besoin ; une typologie qui sera une forme de langage documentaire.

A noter que cette typologie, relativement simple lorsqu'il s'agit de la mettre en place sur un seul support, devra être créée pour prendre en compte plusieurs journaux, avec chacun leurs rubriques et leur organisation.

2 - Recherche d'un type d'articles ou d'un type d'information : portrait, entretien, synthèse, statistiques, chiffres, chronologie

La première caractéristique d'un article est sa taille : une brève, un article long, moyen ou court. Qu'il s'agisse d'une recherche thématique ou d'une question précise, la possibilité de ne pas retenir les brèves permet de limiter le nombre de résultats. On peut bien sûr faire le choix contraire, et privilégier les articles longs où les informations seront plus nombreuses et plus précises. Il sera donc profitable de préciser, brève, long ou court par une surindexation.

Mais cela ne suffit pas. Une des caractéristiques des recherches des journalistes est le besoin d'un type d'articles, d'un type d'informations.
Je cherche un portrait de XXX, la dernière interview de YYY, une synthèse sur tel sujet, le rapport Untel, etc.
Ce type de recherche nécessite d'ajouter aux articles une caractéristique assez subjective qui ne peut être déduite automatiquement de son contenu. Pour cela une indexation humaine est obligatoire.

Si l'on peut caractériser le type d'articles recherchés, on peut affiner la recherche, diminuer le bruit à l'interrogation et augmenter la pertinence des résultats.

De même pour ce qui est de retrouver un type d'information. Les journalistes ont souvent à vérifier un chiffre ou une date, à retrouver un article de loi. Si l'on peut limiter la recherche à des articles contenant des chronologies ou à des tableaux chiffrés, les résultats seront bien meilleurs.

Or un article qui contient le mot chronologie ne sera pas forcément une chronologie. Les statistiques ou les chiffres, lorsqu'ils sont très présents dans un papier, lui donnent une valeur particulière pour ce qui est de la recherche d'informations, mais ne peuvent être repérés par une indexation automatique.

A noter que se pose ici le problème des infographies, très présentes dans les journaux économiques notamment. Pour exploiter ce type d'informations, généralement sous format image, l'indexation humaine est indispensable.

Que ce soit pour caractériser les types d'articles ou les types d'informations, il faudra créer une ou des typologies et donc intégrer un langage documentaire et une indexation humaine. Ces champs, accessibles à la recherche, permettront d'affiner la requête et de diminuer le bruit.

C'est d'ailleurs la solution employée au journal le Monde, avec une typologie par catégories qui a été créée en intégrant notamment les mots outils[6] du langage documentaire. Cette typologie comprend plus de 50 termes et peut être enrichie en fonction des besoins. *(Voir annexe 3 p°V à XI)*
Chez Bayard, chaque dossier documentaire comporte un dossier piste qui rassemble des données historiques, chronologiques ou chiffrées sur un sujet, et qui permet de retrouver rapidement ce type d'informations.
(Voir annexe 2 p°II à IV)

3 - Le langage documentaire permet la structuration des documents et le classement des informations : indexation fonctionnelle et recherche sur le titre

Cette indexation fonctionnelle regroupe en fait une structuration physique objective et une structuration plus subjective qui permet de prendre en compte :

- la place de l'article dans le journal : typologie de rubriques
- le genre de l'article : synthèse, entretien, déclarations
- les caractéristiques de l'information contenue dans l'article

Cette structuration est fondamentale car elle donne par ces typologies une valeur informative à l'article et aux types d'informations qu'il contient.
Cette forme d'indexation permet de sélectionner un corpus d'articles pertinents et d'appliquer ensuite une recherche en texte intégral. Intégré de cette façon, le langage documentaire est complémentaire d'une indexation automatique.
C'est une manière de transposer la médiation documentaire, qui par un dialogue avec l'usager, tente toujours de préciser la demande et notamment le genre de documents et le type d'information recherchés.

La structuration des articles de presse donne accès à un autre type de recherche : la recherche sur le titre, ou plutôt ce que l'on appelle la titraille c'est-à-dire le titre et le chapeau de l'article. Dans la pratique, près de 75% des recherches factuelles se font dans le champ titre.
[26-Journées d'études de l'ADBS]

[6] Dans la presse les mots outils ne correspondent pas complètement à la définition classique du thésaurus ; ils intègrent en effet des termes permettant de caractériser les types d'informations contenues dans les articles (chronologie, statistiques, etc.) et les types d'articles (entretien, synthèse, etc.)

En effet, on peut penser que le titre d'un article contient les informations essentielles pour répondre à une question précise ou factuelle.

De plus, une recherche automatique sur le titre sera moins « bruyante » que sur l'ensemble du texte ; elle pourra donc améliorer ce type de recherches d'informations.

Mais les titres des articles de presse sont rarement parlants et même souvent ambigus. Les titres de Libération par exemple sont bien connus pour être sibyllins.
Si l'on veut mettre en place une recherche efficace sur le titre, cela suppose de l'enrichir manuellement. Cet enrichissement devra rester assez simple et léger.

Il suffira, en fonction des manques observés, de préciser en quelques mots les personnes morales et physiques et les lieux : rajout du prénom sur les noms propres, du nom du pays, précision d'un fait, etc.

On pourra ainsi mettre en place une stratégie de recherche : recherche précise sur le titre et plus générale et thématique sur l'ensemble du texte intégral.
L'indexation humaine et la structuration des documents viendront en complément de la recherche en texte intégral.

C - Besoins d'une synthèse sur un thème, d'un état de l'art sur une question : une indexation thématique et analytique

Nous arrivons là au problème central de la recherche d'informations et de l'indexation. Car ce type de recherche thématique implique une indexation sur le sujet du document, une analyse de son contenu, une réflexion sur son sens, un accès à la connaissance.

Une question sur l'Islam en France aujourd'hui, nécessite de décrypter et d'analyser le sens et les implications du terme Islam, de voir tous les thèmes adjacents, etc.
Une requête qui vise à faire le point sur la crise de la vache folle, oblige à prendre en compte les conséquences au niveau agricole, commercial, politique, santé, etc.

C'est une recherche approfondie, généralement prise en charge par les documentalistes. Pour être efficace, la recherche devra s'effectuer sur le sujet et l'ensemble des thèmes connexes.
L'indexation devra définir, par une analyse du contexte, le ou les concepts composants le sujet, prendre en compte les relations qui les unissent, les thématiques auxquels ils appartiennent.

Ce type de besoins est d'autant plus complexe à satisfaire que les questions sont larges et transversales.
Une recherche thématique se fera le plus souvent sur le texte intégral ou sur les champs sujets de l'indexation humaine. Là aussi, une typologie des articles peut permettre l'affinage de la question. Pouvoir retrouver facilement un article de synthèse est particulièrement important dans le cadre d'une recherche thématique.

Mais c'est avant tout les possibilités d'analyse sémantique et d'organisation des connaissances offertes par la technologie que nous allons tenter de comparer avec les atouts de l'intégration d'un langage documentaire.

1 - La pertinence des documents : niveau d'importance de l'information.

« Le problème de l'accès pertinent à l'information est le problème de l'accès au sens » [17-Théret, p°143]

Dans ce type de recherche une indexation automatique accompagnée de traitements linguistiques ne peut pas suffire.
Les techniques linguistiques limitent certes les ambiguïtés de sens des termes, recherche sur les dérivés (islam, islamique, islamiste) et reconnaissance des expressions (vache folle). Mais la recherche s'effectue sur les mots, le calcul de la pertinence des documents étant déterminé par les cooccurrences de mots entre la requête et les documents.

Seul l'ajout d'une analyse sémantique permet de définir une pertinence par rapport au sens.

C'est grâce à une analyse sémantique que l'on pourra définir réellement la thématique-sujet des documents et de la question.

À chaque mot de la langue, on pourra associer ses différents sens (synonymie, relations d'association et d'appartenance). Pour une recherche sur l'islam en France, une recherche basée sur la sémantique permettra de « définir » l'Islam comme une religion, de retrouver des articles sur les musulmans en France, sur les différents courants de l'Islam, etc.

Une recherche par le sens ne s'emploie pas à retrouver séparément dans un document les éléments de la question. Au contraire, elle va chercher par rapport au sens global de la requête, et le rapprocher du sens global de chaque document. On entend ici par « sens global », une image de l'ensemble des thèmes et sujets abordés dans le document, pondérés par leur fréquence d'apparition.
L'importance d'un document est évaluée en fonction de son sens, par rapprochement sémantique entre la requête et l'article. La qualité de la recherche en est bien sûre améliorée.

Certains moteurs de recherche très développés proposent même plusieurs niveaux de pertinence : pertinence par les mots, par le sens ou un mélange des deux. Le seuil de pertinence est parfois paramétrable à la recherche ce qui permet d'élargir ou d'affiner la recherche. C'est le cas au centre de documentation du journal le Monde où le moteur de recherche Intuition propose 4 niveaux de pertinence, paramétrable à la recherche : 40%, 60%, 80% ou aucune. On peut donc choisir soit d'enlever le tri par pertinence (aucune), soit d'activer une combinaison d'algorithmes propre à Intuition. Ces algorithmes permettent plusieurs niveaux de pondération (40,60 ou 80%) en « jouant » sur les occurrences, la fréquence, la proximité, etc.

La plupart des logiciels s'appuient sur diverses ressources linguistiques qui tendent à représenter les relations sémantiques de la langue en général ou d'un domaine en particulier : réseau sémantique, dictionnaire terminologique ou thésaurus.

La qualité de la recherche dépend de ces outils de représentation sémantique sur lesquels s'appuie l'analyse.

Il est possible d'intégrer des thésaurus ou des dictionnaires terminologiques et de les utiliser comme des ressources sémantiques pour l'analyse. Dans ce cas, on se situe dans le cadre d'une intégration d'une forme de langage documentaire.

D'autres types d'outils de représentation sémantique font de plus en plus l'objet de recherches et de travaux : réseaux sémantiques, ontologie, etc.
Ces outils de gestion du sens tendent à aller vers une analyse sémantique de niveau pragmatique. *(Voir annexe 1 p°I)*

Dans le cas des réseaux sémantiques, « à chaque concept est associé des mots ou expressions, avec leur définition et les relations qui les lient sont identifiées et pondérées selon leur rareté ou leur particularité dans la langue ou dans le domaine. » [16-Leloup, p°73]

Dans le cas des ontologies, « il s'agit de représentations formelles d'un domaine de connaissance sous la forme de terminologies dotées de relations sémantiques (non limitées aux relations sémantiques du thésaurus documentaire). Fondées sur la théorie des graphes, elles font l'objet d'une formalisation logique poussée et ajoutent au thésaurus la notion d'inférence. La finalité de cette formalisation est de rendre les raisonnements accessibles aux machines. » [1-Sajus]

Les progrès effectués et à venir en matière de compréhension de texte et d'analyse automatique du contenu sont assez impressionnants : indexation à rôles, annotations sémantiques des documents, extraction automatique d'informations, création de base de connaissances à partir de textes en langage naturel, etc. [13-Poibeau] [19-Lefèvre] [7Prié]

Quelles que soient ces avancées, l'analyse sémantique automatique pouvant être mise en place aujourd'hui n'atteint pas encore le niveau de la pragmatique.
Les difficultés de compréhension de sens restent encore nombreuses : l'implicite, la connotation, la paraphrase, etc. (Voir annexe 1 p°I)
Toutes difficultés du langage naturel qui sont particulièrement présentes dans le discours de presse.

2 - La classification automatique : une organisation des connaissances ?

Dans une recherche thématique, il s'agit aussi d'accéder à la connaissance d'un sujet, c'est-à-dire de prendre en compte l'articulation d'un sujet autour de différents thèmes, de différents concepts. Même très pertinente, une liste de documents ne peut suffire à ce genre de recherche.

Une classification automatique repère les expressions ou les concepts les plus fréquemment associés aux mots qui composent votre requête, regroupe et organise les documents en fonction de ces expressions.
« Ces techniques visent à déterminer une relation de ressemblance floue entre documents » [11-Le Loarer, p°181]

La classification automatique est un traitement qui peut être effectué après une analyse sémantique et un classement par pertinence de sens, pour organiser les résultats de la recherche (Intuition au Monde).

Mais cette technique, statistique et linguistique, peut être employée par des moteurs de recherche qui ne font pas de traitements sémantiques. C'est le cas sur le web du moteur Exalead.

Une classification automatique sera donc plus ou moins pertinente selon les traitements effectués en amont.

Pour une recherche sur l'Islam en France, le moteur de recherche Exalead propose un classement par ces groupes nominaux : Islam de France, Convertis à l'Islam, Culture Islamique, Associations islamiques, Islam politique, Monde islamique, Piliers de l'islam, République fédérale islamique.

Si on développe quelques concepts, on trouve Musulmans de France, Mosquée de Paris, Jeunes musulmans, Islam Modéré, Conception de la laïcité, Frères musulmans, Histoire de l'Islam, La Mecque, etc.

Certains des concepts extraits ne sont pas pertinents pour la recherche : la catégorie « République fédérale islamique » regroupe des textes sur les Comores. Cependant la classification automatique permet une visualisation des thèmes abordés par les documents.

Une classification de ce type est donc plus une aide à la définition du sujet qu'une véritable organisation des connaissances. Le classement par concepts des documents apporte du sens à la question, éclaire les différents thèmes la composant et permet un accès facile pour les utilisateurs non-spécialistes. Cependant elle ne permet pas de situer les documents dans de grands domaines de connaissances. Il n'y a pas réellement de définition de ces classes ou de ces concepts. Si le concept « jeunes musulmans » paraît représenter un thème cohérent, quant est-il de la définition du contenu des textes regroupés autour de l'expression « musulmans de France ». Rien ne nous dit si ce sont des articles politiques, culturels ou religieux.

Ni organisation des connaissances, ni véritable classement, la classification automatique est un outil de représentation du contenu.

Cette représentation du contenu permet en fait une conceptualisation de la question posée en fonction des résultats, et facilite donc la reformulation de la requête. C'est un outil très efficace pour gérer plus facilement le bruit à la recherche, en éliminant les thèmes qui ne conviennent pas.

Pour une véritable organisation des connaissances, il faudra mettre en place une classification automatique s'appuyant sur des catégories préexistantes.
Ces catégories peuvent être établies par exemple à partir des champs sémantiques d'un thésaurus, la classification s'appuie alors sur une forme de langage.
Il est possible également de « ranger des documents dans une ou plusieurs classes définies par des textes de références ou des profils d'intérêts » mais un contrôle manuel reste indispensable. [19-lefevre p°129]

3 - Une indexation humaine conceptuelle en profondeur et un langage documentaire pour ordonner les domaines du savoir

Nous avons vu déjà que l'intégration de tout ou partie d'un langage améliore la qualité de la recherche sémantique ; voyons maintenant ce que peut apporter l'intégration directe d'un langage par indexation humaine.
L'indexation se fera dans un champ sujet ou thèmes, rempli manuellement par les documentalistes à l'aide du langage.

Malgré les progrès constants de la technique, l'indexation humaine permet une finesse d'analyse et une conceptualisation des documents qui reste fondamentale pour la qualité et la pertinence d'une recherche thématique. Et ce malgré les problèmes posés par le manque d'homogénéité de l'indexation manuelle.

Préserver l'indexation humaine c'est préserver la qualité de l'analyse intellectuelle d'un document. [9-Ward]
Cette analyse est d'autant très importante dans le contexte de la presse, étant donné les ambiguïtés du discours et les difficultés de décryptage des articles.

En effet, que ce soit avec l'aide d'un thésaurus ou un index matière, l'indexation manuelle est la seule qui puisse transcrire et conceptualiser des notions qui ne sont pas explicites dans un document.

L'indexeur pourra prendre en compte les connotations et les non-dits d'un article, faire des rapprochements conceptuels grâce à sa connaissance du domaine. Par exemple, on pourra indexer un article sur les dernières négociations de l'OMC[7] par le descripteur relations NORD/SUD. Même si ce terme n'apparaît jamais, les négociations commerciales internationales mettent en jeu ce type de thématique.

Autre avantage, même si une notion est exprimée de manière différente, elle est représentée par le même descripteur normalisé.
La connaissance du domaine et des sources permettra de décrypter les différences de formulation entre les différents journaux. L'extrême gauche du Figaro n'est pas celle de l'Humanité et l'utilisation d'un même descripteur permet une description uniforme des documents.

De plus, une indexation humaine commence par une sélection des articles qui permet de limiter la redondance de l'information de presse.
Il s'agira de privilégier la meilleure source d'information et de garder un article économique de la Tribune ou des Échos, plutôt que celui du Parisien abordant le même sujet.

Mais il s'agit aussi de garder les articles les plus importants en fonction du thème abordé. L'indexeur différenciera dans un article, le concept central de celui qui est simplement cité.

[7] OMC : Organisation Mondiale du Commerce

L'apparition du terme « politique étrangère de la France » ne suffit pas à en faire le sujet du document, le rôle de cette expression dans la phrase peut être anecdotique.

L'indexation permettra une sélection et une indexation fine des articles, et donc une plus grande pertinence des documents à la recherche et une diminution du bruit. La pertinence des documents dépend de la justesse de l'indexation par sujet et permet une recherche plus exacte.

Bien sûr, ce genre d'indexation génère aussi du silence dans la mesure où toutes les notions présentes dans un article ne peuvent être décrites. En effet, si l'indexation est trop profonde (utilisation de nombreux descripteurs), non seulement elle perd de son sens mais elle génère beaucoup de bruit. Une indexation humaine se doit de caractériser un article par ces notions les plus importantes, et deux ou trois concepts sont largement suffisants pour définir une thématique-sujet.

Il s'agit en fait de repenser l'indexation matière en complément du texte intégral, afin d'enrichir la recherche en texte intégral avec une interrogation par mots-clés. [25-Villacampa]
La surindexation devra donc être suffisamment légère pour diminuer le coût et le temps d'indexation, tout en permettant la prise en charge des non-dits et l'organisation des connaissances.

L'indexation humaine à l'aide d'un langage place immédiatement le texte dans un réseau de relations sémantiques qui sont porteuses de sens : relations d'association, d'équivalence ou de hiérarchie, d'appartenance à un champ sémantique. La finesse d'un thésaurus, et de ses liens sémantiques, est particulièrement adaptée à l'indexation thématique et à la définition du sujet.

Prenons un exemple et voyons la conceptualisation et l'organisation du langage chez Bayard.
Les articles sur la canicule de cet été en France étaient tous indexés à météo, mais aussi selon l'angle du papier à hôpital/généralités, troisième âge/santé, etc.
Le descripteur hôpital/généralités appartient au champ sémantique santé et médecine, il est le père d'urgence médicale et est associé au descripteur santé qui regroupe les articles sur le système de santé en France.
Le concept troisième âge/santé est proche de la notion de dépendance et de troisième âge/hébergement (ce qui concerne les maisons de retraite), il appartient au champ société et on peut être renvoyé à famille/relations.

Le choix d'un de ces descripteurs, non seulement défini le contenu du texte mais aussi son appartenance à un domaine. Un article indexé avec hôpital/généralités fait partie du domaine des questions de la santé en France ; un article indexé avec troisième âge/santé traite un thème de société.

C'est cette organisation des connaissances qui permettra une véritable définition du sujet, une conceptualisation par rapport à un contexte et à des savoirs.

D - Besoin de feuilletage et boite à idées : la fonction de navigation par le langage en soutien de l'hypertexte et de la classification

Pratiquement, dans un fonds documentaire, le journaliste consulte des dossiers thématiques papier autour d'un ou plusieurs thèmes afin d'avoir une vision de la façon dont un sujet a déjà été traité ou pour trouver un angle ou une idée originale.

Prenons le cas d'un journaliste intéressé par une recherche autour du thème de la laïcité en France. La plupart du temps, la formulation de base de son besoin ne sera pas plus explicite.
La documentaliste traduit et développe les centres d'intérêts du journaliste et l'oriente vers tel ou tel dossier : la laïcité de l'enseignement, la laïcité des institutions, etc. Le langage documentaire est médiateur dans la mesure où il permet cette traduction et formalise les liens entre les différents thèmes. La lecture papier permet de feuilleter rapidement et de trouver la bonne information ou de préciser sa question.
Même si la recherche est effectuée au final par le journaliste, la médiation documentaire est fondamentale.

C'est, sans aucun doute, le type de recherche le plus difficile à adapter à l'informatique, à reproduire dans une base de données.
Si la base intègre un langage documentaire, la navigation entre les concepts permettra de « piocher » en fonction de ses centres d'intérêts et de formuler une recherche thématique « floue ».
En l'absence d'un langage, c'est par la représentation des résultats et les outils d'aide à la recherche qu'il faudra recréer cette possibilité.

« Plus l'autonomie de l'utilisateur est importante, plus l'outil permettant le développement du questionnement doit être performant. Et cette performance ne consiste pas à éluder la phase d'élaboration de la question mais au contraire à l'approfondir. » [1-Sajus]

1 - L'hypertexte et la classification : une nouvelle approche de la représentation du contenu

Pour ce type de recherche, l'ergonomie de la présentation des résultats, devra permettre la sélection rapide de l'information pertinente, une réorientation de la question et une consultation du document qui soit adaptée.
A tous ces niveaux, des outils spécifiques tendent vers une nouvelle approche de la représentation du contenu.

« La pertinence de la présentation des résultats peut s'évaluer à partir de plusieurs indicateurs, tels les moyens fournis pour apprécier l'adéquation des résultats à la question, ou encore, les moyens pour juger de l'opportunité ou non de consulter l'intégralité de la ressource (étape de sélection de l'information) » [14-Sébillot, p°48]

Lors de la consultation d'un dossier documentaire, on sait très vite si un article va être intéressant par une lecture transversale du titre, chapeau, paragraphe, etc.
L'affichage d'un document devra donc être accompagné du titre et du chapeau, du surlignage des groupes nominaux comprenant au moins un mot important de la requête, de son niveau de pertinence, de sa taille, etc. L'ensemble de ces informations permettra la sélection des articles intéressants.

Viennent ensuite les fonctions de précision, reformulation ou affinage de la question. Dans une liste de réponses, le fait de qualifier comme bons ou mauvais les documents au gré des consultations, permet d'apprendre au système vos attentes réelles. La fonction d'affinage permet alors de reposer la requête correspondante, afin de rectifier la pertinence de chaque réponse.

La recherche par l'exemple peut également être utile. Cette fonction de navigation particulièrement intéressante consiste à poser en requête la totalité d'un document, que l'on juge pertinent, et qui nous a été proposé dans une liste de réponses antérieures.

Mais il arrive fréquemment que la liste de réponses obtenue soit trop longue pour son exploration exhaustive. Cela arrive d'autant plus fréquemment que la question est vaste ou ambiguë.

La classification automatique offre alors la possibilité de sélectionner aisément des sous-ensembles de réponses plus proches des attentes de l'utilisateur.
Il est possible d'affiner sa requête par un simple « clic » sur le concept jugé le plus pertinent, ou au contraire d'éliminer un groupe de documents non pertinents. Il est donc possible de réorienter la recherche en croisant la requête avec un des concepts.

Mais une navigation dans ces concepts est également possible grâce à des liens hypertextuels, et c'est cette navigation qui tente de reproduire le phénomène de feuilletage.

La classification peut également être intéressante pour la détection des signaux faibles. En effet, une classification permet de détecter les expressions qui apparaissent le plus dans les documents mais aussi les concepts qui apparaissent le moins. On peut ainsi se faire une idée des angles les moins traités dans les articles, des associations les plus originales.

Les logiciels de cartographie permettent le même genre de « recherche floue », avec en plus une visualisation de l'ensemble des thèmes de la question, de leur importance et de leurs liens.

Classification ou cartographie, la navigation permet à l'utilisateur de sélectionner les éléments qu'il juge utiles et de spécifier les relations entre ces éléments pour former une requête documentaire complexe.
[8-Médini et Bigeon]

Enfin vient la consultation du document. Les bases de données de presse proposent la lecture des articles en format PDF afin de donner la vraie mise en page de l'article dans le journal et de tenter de reproduire, par une mise en image, la lecture papier.

A tous ces niveaux, la classification et la navigation hypertexte se posent en médiateur entre l'utilisateur et le système : conceptualisation de la question, reformulation, aide à la recherche, navigation, etc.
L'ensemble de ces techniques vise à replacer l'utilisateur et son questionnement au centre des systèmes d'information. Au cœur de ces systèmes, les outils de représentation du contenu et de navigation tendent à mettre en place un lien, une médiation entre l'utilisateur et l'information.

Malgré tous ces progrès, la lecture papier déclenche des processus cognitifs qui, à mon avis, ne peuvent pas être reproduits sur écran. Le feuilletage d'un dossier documentaire reste un processus difficilement adaptable.

2 - Le langage documentaire : un médiateur pour la recherche, un lien avec les utilisateurs

Pour ce type de recherche floue, l'intégration d'un langage documentaire est utile à plusieurs niveaux.

Nous avons vu que l'indexation humaine permet de prendre en compte les connotations d'un document.
L'indexation de connotations permet de définir un texte par rapport à des concepts abstraits comme l'amour ou l'amitié, et ouvre la voie à des recherches plus larges, plus transversales. La notion d'amour pourra regrouper à la fois des articles psychologiques sur le sentiment amoureux, mais aussi des papiers sur la sexualité, sur les rapports homme/femme, sur le mariage et le concubinage, etc.
L'interrogation sera facilitée car elle pourra se faire à l'aide d'un seul concept.

Quelle que soit la qualité de la classification et de l'analyse sémantique, elle ne peut détecter entre les articles ce type de parenté invisible.
L'indexation humaine peut, à ce niveau, prendre en compte les centres d'intérêts des utilisateurs, intégrer de nouvelles notions en fonction de leurs besoins.

L'intégration d'un plan de classement ou d'un thésaurus permet, avant même l'interrogation, de se faire une première idée du contenu de la base de données. La consultation du langage permet de savoir quels sont les thèmes abordés, de quelle façon ils sont reliés entre eux, comment les connaissances sont organisées. Par exemple, le langage documentaire chez Bayard permet de prendre connaissance des différentes subdivisions par pays et de leur contenu : culture, économie, défense, environnement, etc. (Voir annexe 2 p°II à IV)

Le langage se pose donc en préalable à la recherche : si on sait à quoi on peut avoir accès, on peut alors chercher.

Mais le langage aide surtout à lever une des difficultés les plus importantes de la recherche, la conceptualisation et la formulation de la question.
[20-Collas et Chartron]

Le langage se pose alors comme un outil intellectuel d'aide à l'interrogation.
Il est une aide au questionnement dans la mesure où il propose l'explicitation et la formalisation des relations sémantiques qui unissent les termes d'un domaine. A ce niveau préalable, la navigation dans un thésaurus ou dans un plan de classement permet de se faire une idée de l'organisation d'un domaine, de trouver des idées, de découvrir un angle de vue.

Il permet par exemple, non seulement de savoir qu'il existe des articles sur les femmes battues, mais aussi d'apprendre, par la relation d'association, qu'il en existe d'autres sur les hommes battus.

Reprenons l'exemple de la question sur l'islam en France et voyons en quoi le langage documentaire de Bayard peut aider à formuler la question.
Le thème de l'islam appartient au champ religion, avec comme environnement sémantique : catholicisme, christianisme et judaïsme. Plusieurs descripteurs organisent les différents thèmes : ISLAM/FRANCE, CATHOLIQUE/MUSULMAN, INTEGRISME/MUSULMAN, ISLAM/GENERALITES.
Le contenu des concepts est précisé : islam/France rassemble notamment les associations et les représentations musulmanes, Islam/généralités contient l'histoire et les doctrines de l'Islam.
D'autres descripteurs sont associés (voir aussi) : jeune/religion, immigration.
Des renvois ont été crées : les problèmes du port du foulard islamique seront indexés à laïcité et tout ce qui concerne l'islam en dehors de France sera à Pays/Religion.

L'organisation des connaissances, autour de ce thème, est précise et définie. La recherche sera guidée par ce réseau de liens. Par la consultation du langage on pourra avoir une idée globale de l'organisation d'un thème et donc préciser ou élargir sa recherche.

La navigation permet à l'utilisateur de passer en revue de nombreux éléments qui peuvent être pertinents pour son besoin d'information.

Techniquement, la navigation dans un langage est facilitée par l'utilisation de liens hypertextes. Au moment de l'interrogation, l'autopostage permet une aide à la reformulation de la question. L'autopostage permet en effet d'élargir automatiquement la recherche à des descripteurs reliés par des liens sémantiques. On pourra donc élargir la recherche en prenant en compte les descripteurs plus larges ; préciser la question en reformulant par un descripteur plus précis ; transformer l'interrogation par l'utilisation de descripteurs associés.

« L'utilisateur a besoin d'un espace intermédiaire, essentiellement un espace de mots qui lui permettra de partir de ce qu'il connaît déjà pour tracer un chemin vers ce qu'il ne connaît pas encore. » [1-Sajus]

La médiation d'un langage permet de réintroduire un lien, un dialogue avec l'utilisateur lors de la phase de questionnement et d'interrogation. La fonction du langage dans une base de données est bien celle de la médiation documentaire.

E - Autres fonctions du langage

Nous avons vu les performances, les avantages et les inconvénients de l'utilisation d'un langage documentaire ou de l'indexation automatique au niveau de la recherche d'informations.
Cependant, certaines fonctions du langage, qui me paraissent importantes également, ne sont pas directement liées à la problématique de la recherche documentaire.
Sans prétendre être exhaustive, je voudrais souligner trois des fonctions de l'utilisation d'un langage documentaire dans un centre de documentation. Ces fonctions persistent si l'on choisit d'intégrer le langage à un système automatique d'indexation automatique et de préserver une part d'indexation humaine.

1 - La préservation de l'identité et de la culture

Utiliser un langage dans un centre de documentation, signifie le construire et le développer en fonction, non seulement du domaine de connaissances mais aussi par rapport aux besoins des utilisateurs. En somme le développement d'un langage doit pouvoir représenter l'identité de l'entreprise. Le langage est porteur de connaissance de la culture de l'entreprise. Cet outil doit être l'objet d'une constante surveillance pour l'actualiser, l'entretenir et l'alimenter.
[25-Villacampa]

Chez Bayard, le langage prend en compte les domaines spécialisés traités par les publications et donc les publics particuliers auxquels elles s'adressent. Les journaux du groupe sont soit centrés sur le thème de la religion catholique, soit s'adressent spécialement à la jeunesse ou aux seniors.
Les questions des journalistes sont donc particulièrement nombreuses et précises sur ces trois thèmes.
De plus, Bayard étant un groupe de presse, tout ce qui a à voir avec les médias est privilégié, de la presse écrite à Internet en passant par l'audiovisuel.

Le langage documentaire s'est donc enrichi en fonction de l'importance de ces questions et présente aujourd'hui un niveau de complexité et de finesse plus important sur ces thèmes. Les descripteurs ayant trait à la religion sont plus nombreux et plus précis. Le « champ médias » est particulièrement développé. *(Voir annexe 2 p°II à IV)*

Le langage documentaire s'est développé pour mieux répondre aux utilisateurs, pour mieux « coller » à la culture de l'entreprise.
Préserver le langage c'est aussi préserver une part de l'identité de l'entreprise, une part de sa richesse.

2 - Le lien avec le fonds papier

L'intégration d'un langage dans le cadre de l'informatisation d'un service de documentation, c'est aussi permettre de garder le lien avec le fonds papier.

Même si dans le cadre de la coopération documentaire, les hebdomadaires et les quotidiens français sont accessibles sous format numérique, c'est loin d'être le cas pour les autres sources des centres de documentation.
Et il faudra sans doute attendre encore longtemps pour que l'ensemble des revues spécialisées soit disponible sous une forme adéquate. Le dépouillement papier sera encore longtemps utile.

Pour pouvoir faire une recherche à la fois sur une base de données et sur ce qui reste en papier il faut absolument préserver soit un langage d'indexation commun, soit une passerelle sous la forme d'une liste de renvois. Il faut pouvoir « interroger » avec les mêmes termes et garder une analyse et un classement uniforme.

Les centres de documentation possèdent souvent une bibliothèque où les livres utiles à la recherche doivent être retrouvés facilement. Si l'on peut intégrer les notices de ces ouvrages dans une base de données, il faudra qu'ils aient été indexés avec un langage similaire pour les retrouver lors d'une recherche.

3 - L'indexation comme mode de connaissance d'un domaine

Enfin, un problème connexe. Si on fait disparaître complètement l'indexation humaine, qu'elle soit au niveau informatique ou papier, on fera disparaître dans le même temps, la connaissance que peut apporter cet exercice.

En effet, devoir sélectionner et analyser les articles de presse, les dépouiller chaque jour, comprendre le thème central de chaque document, trouver le descripteur adéquat, etc. Tout cela apporte une connaissance particulière qui permet ensuite de mieux maîtriser le domaine de connaissance couvert.
Lorsqu'on a indexé un document soi-même, il est beaucoup plus facile de le retrouver. La pratique de l'organisation d'un langage permet d'aider plus facilement à reformuler une question d'un utilisateur.
La pratique de l'indexation est donc un gage de compétences pour les documentalistes, un gage de qualité pour la médiation documentaire.
Sans un langage et sans la pratique de l'indexation, il faudra développer d'autres techniques d'apprentissage du domaine de connaissances.

CONCLUSION : L'AVENIR DES LANGAGES DOCUMENTAIRES, UN ENJEU DE L'ACCÈS À L'INFORMATION

Il s'agit maintenant de tirer un bilan sur la place des langages documentaires dans les systèmes d'information presse face aux nouvelles technologies.
L'expansion de l'interrogation en texte intégral et les progrès du TALN ont pour corollaire une diminution de l'indexation humaine et des langages documentaires qui la formalisent.
Ce phénomène de fond est une réalité qu'il ne faut pas nier. Mais dans le contexte de la presse, nous avons vu combien l'intégration d'un langage documentaire améliore la qualité de la recherche.

Persistance de certaines fonctions du langage documentaire

Les progrès de l'ingénierie linguistique parviennent à gommer une partie des inconvénients du langage naturel, et la classification automatique facilite la gestion du bruit persistant.
Mais dans la presse, les ambiguïtés et l'évolution du discours nécessitent la normalisation d'un langage. L'indexation automatique restera de qualité médiocre si on ne lui adjoint pas au moins un dictionnaire de synonymes…embryon d'un langage documentaire.

De même la fonction de médiation d'un langage documentaire à la recherche reste fondamentale. Plan de classement des documents d'une base de données, outil d'organisation conceptuelle d'un domaine, de questionnement et de navigation entre les sujets, les langages documentaires restent indispensables pour le rapprochement linguistique et sémantique entre la question de l'usager et le contenu d'une base de données. La fonction heuristique du thésaurus est une aide à la recherche sur le texte intégral et vient en complément des classifications et de la navigation hypertexte.

Changement dans l'indexation : un usage sélectif et adapté

Les performances des outils de TALN et de recherche remettent en cause une partie, non pas du langage mais de l'indexation humaine.

En réalité il s'agit de faire un usage sélectif de l'indexation humaine, en fonction des besoins des utilisateurs et en vue d'enrichir la recherche en texte intégral.
Nous avons vu l'importance de l'indexation fonctionnelle et de la structuration des documents. Les typologies de documents permettent d'adapter la recherche aux besoins des utilisateurs et de redonner des niveaux d'importance aux types d'informations.

Dans le contexte de la presse, l'indexation humaine reste très utile pour donner du sens, à la fois dans la prise en compte de la connotation des mots, des concepts abstraits et des parentés invisibles entre documents.
Indexer humainement c'est aussi préserver la richesse que peut apporter la connaissance du domaine et des besoins des utilisateurs.

Évolution des langages documentaires : vers l'organisation des savoirs

Intégrés dans des systèmes de classification automatique, ou comme ressources pour l'analyse linguistique sémantique, les langages documentaires deviennent de plus en plus invisibles.
Si leur forme évolue pour s'adapter à la technologie, leur rôle d'outil sémantique est préservé. Ils apportent de la spécificité et de la précision là où les dictionnaires restent très généraux.
C'est la finesse de leurs liens sémantiques, mais aussi leur capacité à formaliser un domaine de connaissance qui restent indispensables.
S'ils doivent se transformer pour mieux s'adapter aux outils informatiques, leurs fonctions de base au niveau de la gestion du sens doivent être préservées.

Mais il est sans doute important de repositionner les langages documentaires, et spécialement le thésaurus par rapport aux nouveaux outils et méthodes de gestion sémantique comme les ontologies. Il est également fondamental que les documentalistes participent à la normalisation de ces nouveaux outils.
[1-Sajus]

Les conséquences de l'évolution de la place des langages dans les systèmes d'information

Bien sûr, l'évolution des langages documentaires, de leur place dans les systèmes d'information provoquent également des changements dans le métier de documentaliste.

Pour les documentalistes, cela signifie d'abord une diminution du temps d'indexation et donc plus de temps pour des taches à plus forte valeur ajoutée comme les recherches thématiques précises.
Le temps gagné dans le traitement du document devra être reporté sur un travail d'analyse en amont. Cela signifie donc une plus grande réflexion sur le type d'indexation utile aux utilisateurs, conforme à leurs pratiques et qui serait complémentaire de recherche en texte intégral.
[24-Dalbin et Saleras, p°323-324]

Cela veut dire s'adapter à d'autres types de recherche, d'autres modes de représentation du contenu, générés par les classifications par exemple, et participer à l'élaboration de l'ergonomie des bases de données.

Enfin, le travail sur les traitements linguistiques n'est pas de même nature que celui lié à la maintenance d'un thésaurus. Il s'agira donc de se former pour pouvoir travailler sur des dictionnaires spécialisés, des ontologies, etc.
[24-Dalbin et Saleras, p°323-324]

De même, il est important, pour les documentalistes, de connaître la base de la structuration XML, qui est adopté aujourd'hui dans de plus en plus de systèmes d'information.

L'accès à l'information : un enjeu particulier

La recherche sur le texte intégral devient aujourd'hui le mode d'accès de base à l'information dans de plus en plus de domaines. Il n'y a plus aujourd'hui d'opposition entre les outils technologiques et les langages mais des expérimentations de collaboration et une complémentarité certaine. Pour une recherche efficace, il est indispensable d'intégrer à la fois les techniques linguistiques, sémantiques et statistiques, et des techniques documentaires basées sur le langage et l'indexation.

Mais la question se pose aujourd'hui de savoir qui aura accès à ces systèmes d'information sophistiqués, à qui sera réservé ce mode d'accès à l'information ?
En effet, ces systèmes, dont le coût de mise en place est important, paraissent être pour l'instant réservés aux professionnels, comme des produits de luxe à une clientèle privilégiée.
Dans cette perspective, le web et les cédéroms ne seraient que des produits grands publics ne donnant qu'un accès relativement médiocre à l'information : peu de technologies linguistiques et sémantiques, pas de langage et de classement, peu de structuration, etc.

Car c'est bien la maîtrise du langage, même intégrée dans les traitements linguistiques et sémantiques, qui est l'enjeu de cet accès à l'information.
Les réflexions menées autour du web sémantique reposent sur cette problématique et s'il voit le jour, il permettrait de donner, grâce aux langages, un meilleur accès à l'information.
[30-Archimag] [31-Rasmussen] [32-Adams]

BIBLIOGRAPHIE THEMATIQUE

Les références de la bibliographie sont classées en grands thèmes qui sont les grands axes de ma problématique.
Le premier thème regroupe les questions centrées sur les langages documentaires, leurs différentes formes et leurs évolutions possibles.
Le second point aborde les problèmes liés à l'indexation et les interrogations autour de l'indexation automatique.
En troisième lieu sont regroupés les analyses sur le traitement automatique du langage naturel (TALN) et les outils linguistiques qui en découlent.
Un quatrième thème est plus centré sur la recherche documentaire et sur les questions de recherche en texte intégral.
La cinquième partie permet d'illustrer l'ensemble par des exemples concrets de mise en place de systèmes d'information documentaires.
Le dernier point illustre les nouvelles problématiques sur le web sémantique.

Les ouvrages sont classés par thème, puis par ordre chronologique (du plus récent au plus ancien) à l'intérieur de ces thèmes.
Les références faites aux différents textes dans le corps du mémoire sont normalisées sous la forme d'une numérotation qui suit l'ordre de la bibliographie. Pour une meilleure lisibilité, j'ai ajouté le nom de l'auteur du document à ce numéro, et la page de référence lorsqu'il s'agit d'une citation.

Tous les textes de la bibliographie ne seront pas cités dans le cours du mémoire, dans la mesure où certains m'ont simplement servi à prendre connaissances d'un domaine ou à découvrir une question connexe à ma recherche.

I/ Langages documentaires (4 références)

[1-Sajus]
La fonction thésaurale au cœur des systèmes d'information [ressources électroniques] / Bertrand Sajus. In Journées d'étude ADBS : Du thésaurus au web sémantique, les langages documentaires ont-ils encore un avenir - 11 avril 2002. [consulté le 10 juin 2003] .-
<http://www.adbs.fr/uploads/journees/1082_fr.php>
Il s'agit ici de mettre en évidence l'actualité du concept de thésaurus face à l'évolution technologique des systèmes d'information et l'émergence du web sémantique.

[2-Maniez]
Actualité des langages documentaires : fondements théoriques de la recherche d'informations / Jacques Maniez.- Paris : ADBS éditions, 2002.- (Collection Sciences de l'information, série Études et Techniques).- 396 p.- ISBN 2-84365-060-7.- ISSN 1160-2376
L'objet de cet ouvrage est de préciser le rôle et l'importance que conservent les langages d'indexation et de recherche à l'aube du XXIe siècle. J'ai particulièrement utilisé le chapitre VII sur la problématique des langages documentaires

[3-Barite]
The notion of "category": its implications in subject analysis and in the construction and evaluation of indexing languages / Mario Guido Barite.- Knowledge organization, 2000, vol.27, n°1-2, p.4-10
La notion de catégorie, depuis Aristote, Kant jusqu'au temps présent, a été employée comme un outil intellectuel de base pour l'analyse de l'existence. Cet article propose un réexamen conceptuel et méthodologique de la notion de catégorie, dans une perspective fonctionnelle et instrumentale, et essaye de clarifier les caractères essentiels des catégories, et leurs implications quant à la construction et l'évaluation de langages d'indexation.

[4-Maniez]
L'évolution des langages documentaires / Jacques Maniez.- Documentaliste-Sciences de l'information, juillet-octobre 1993, vol.30, n°4-5, p.254-259
Maniez étudie le développement des deux grandes familles de langages documentaires (classifications hiérarchiques et langages analytiques): il cherche à déterminer de quelle manière les éléments des systèmes de recherche d'informations - documents, indexeurs, concepteurs et usagers, techniques de recherche, techniques d'indexation ont influencé cette évolution.

II/ Indexation et indexation automatique (7 références)

[5-Petit]
Améliorer la performance des langages d'indexation : travail sur les index et le thésaurus du centre de documentation des arts et métiers / Violette Petit.- 2002.- 135 p. Mémoire DESS : Info-doc. : Paris, INTD/CNAM : 2002
Ce mémoire s'appuie sur un travail effectué sur l'amélioration des langages documentaires au Musée des arts et métiers. Il tente de définir les particularités des langages d'indexation, leurs méthodes de construction et les moyens mis à disposition pour juger de la performance d'un langage, tant à l'indexation qu'à la recherche.

[6-Amar]
Les fondements théoriques de l'indexation : une approche linguistique / Muriel Amar.- Paris : ADBS éditions, 2000.- (Collection Sciences de l'information, série Recherches et documents).- 355 p.- ISBN 2-84365-042-9.- ISSN 1159-7666
Cet ouvrage tente de dégager les spécificités, les caractéristiques, les propriétés de l'indexation et, pour ce faire, d'en étudier les fondements théoriques par le biais d'interrogations issues de la linguistique.

[7Prié]
Sur la piste de l'indexation conceptuelle de documents : une approche par l'annotation / Yannick Prié.- Document numérique [numéro spécial sur l'indexation], 2000, vol 4, n°1-2, p°11-35
L'indexation conceptuelle est ici définie comme l'ensemble des connaissances ajoutée à un document et pouvant servir à son exploitation, pourvu que cette connaissance soit utilisable aussi bien par l'homme que par la machine.
Un article intéressant sur l'annotation sémantique et la structuration XML.

[8-Médini et Bigeon]
Intégration de l'indexation conceptuelle dans l'expression du besoin d'information / Lionel Médini et Philippe Bigeon.- Document numérique [numéro spécial sur l'indexation], 2000, vol 4, n°1-2, p°85-108
L'auteur développe les différentes stratégies d'accès à l'information en fonction des besoins de l'utilisateur : aide à la formulation, navigation, diffusion sélective de l'information. De l'utilisateur comme dimension d'accès à l'information.

[9-Ward]
The Future of the human indexer / Martin L. Ward.- Journal of Librarianship and Information Science, December 1996, vol.28, n°4, p.217-225
Les besoins de l'indexation et des aptitudes intellectuelles qu'elle requiert sont examinés afin de déterminer ce que devrait être un système automatique d'indexation pour remplacer ou compléter un indexeur. Une bonne indexation, notamment dans les domaines techniques, exige : une connaissance préalable considérable de la littérature ; un jugement de valeur du document, de l'utilisation de l'index, de la profondeur de l'indexation ; des qualités de lecture ; des aptitudes à résumer, cataloguer et classifier.

[10-Fidel]
User-Centered Indexing / Raya Fidel.- Journal of ASIS, September 1994, vol.45, n°8, p.572-586
Deux approches différentes du processus de l'indexation sont analysées : l'approche orientée sur le document qui tend à représenter son contenu et l'approche centrée sur la requête de l'utilisateur. Cette dernière approche, moins classique, mériterait d'être plus développée. Elle est en effet mieux adaptée aux exigences de l'indexation automatique mais nécessite de bien comprendre les utilisateurs et leurs comportements de recherche.

[11-Le Loarer]
Indexation automatique, recherche d'information et évaluation / Pierre Le Loarer. In : Le traitement électronique du document / Cours INRIA 3-7 octobre 1994, Aix en Provence.- Paris : ADBS Éditions, 1994.- p.149-201.- ISBN 2-901046-76-2
Cet article tente de définir les différents types d'indexation : indexation automatique, indexation humaine, indexation libre. Puis il propose une typologie de ces différents modèles d'indexation : indexation structurée, indexation pondérée,... Il explique les techniques d'indexation automatique, quelles soient linguistiques ou statistiques. Enfin il tente d'évaluer les avantages et les inconvénients de chacune de ces méthodes

III/ Traitement automatique du langage et Outils linguistiques (6 références)

[12-Chaumier et Dejean]
Recherche et analyse de l'information textuelle : tendances des outils linguistiques / Jacques Chaumier et Martine Dejean.- Documentaliste-Sciences de l'information, février 2003, vol.40, n°1, p.14-24
Fondée sur une enquête menée en 2002, cette étude propose une analyse de l'offre actuelle d'outils de recherche et d'analyse d'information textuelle. Les auteurs exposent en détail les principes de l'approche linguistique qui préside à la conception de la plupart des logiciels d'indexation assistée par ordinateur, puis esquissent une typologie des outils linguistiques existants et une autre des applications documentaires faisant appel à ces outils.

[13-Poibeau]
Extraction automatique d'information, du texte brut au web sémantique / Thierry Poibeau.- Paris : Hermès, 2003.- 238 p.- ISBN 2-7462-0610-2
Cet ouvrage présente les progrès récents en extraction d'information et en compréhension de textes. Les recherches effectuées ces dernières années dans le domaine du traitement automatique des langues rendent en effet possible l'annotation sémantique des documents, l'extraction d'information pertinente et la création de bases de connaissances structurées à partir de textes en langage naturel.

[14-Sébillot]
Traitement automatique des langues et recherche d'information / Pascal Sebillot. In : La recherche d'information sur les réseaux / Cours INRIA 30 septembre-4 octobre 2002, Le Bono.- Paris : ADBS Éditions, 2002.- p.137-168.- ISBN 2-84365-062-3
L'objectif de ce chapitre est de proposer un tour d'horizon des méthodes, ressources et outils issus du traitement automatique des langues et qui peuvent s'intégrer dans des systèmes de recherche d'information et améliorer les capacités des moteurs de recherche.

[15-Chaumier et Dejean]
Le rôle des techniques linguistiques pour la recherche documentaire / Jacques Chaumier, Martine Dejean.- Document numérique, juin 1997, vol.1, n°2, p.169-176
La recherche documentaire dans les bases de données textuelles passe par des processus mettant en jeu divers types d'approches linguistiques : recherche par mot-clef, recherche avec les méthodes dites de recherche en texte intégral, recherche en langage naturel. Ce dernier point requiert divers niveaux d'analyse : analyse morphologique, analyse syntaxique, analyse sémantique.

[16-Leloup]
Moteurs d'indexation et de recherche : environnement client-serveur Internet et Intranet / Catherine Leloup.- Paris : Eyrolles, 1997.- 285 p.-ISBN2-212-08976-7
Présentation des outils permettant d'accéder rapidement à l'information grâce à la recherche sur le contenu des documents textuels. Le livre explique comment fonctionne ces outils et comment les intégrer aux systèmes documentaires, aux bases de données ou aux serveurs web.

[17-Théret]
Besoin en traitements automatiques du langage naturel pour la recherche d'information sur les réseaux / Philippe Théret. In : La recherche d'information sur les réseaux, Internet pour en savoir plus / Cours INRIA 30 septembre-4 octobre 1996, Trégastel.- Paris : ADBS Éditions, 1996.- p.127-164.- ISBN 2-901046-62-2
Cet article traite du problème de l'accès à l'information sur Internet et de la pertinence de cette information. Trois grandes phases sont détaillées : indexation des documents sources, traitement de la requête et application de la mesure de pertinence. Après un bilan plutôt mitigé, il définit les objectifs à atteindre. Puis il explique les différentes couches des traitements automatiques du langage naturel et tente d'en mesurer l'impact sur la pertinence.

IV/ Recherche d'informations (5 références)

[18-Blair]
Language and representation : 1)Information retrieval and the philosophy of language / David C Blair.- Annual review of information science and technology, 2003, vol.37, p.3-50
La recherche documentaire est fondamentalement un processus linguistique : nous décrivons ce que nous recherchons et tentons de faire correspondre cette description avec les descriptions de l'information qui nous sont disponibles. L'auteur se concentre sur une des activités les plus problématiques de la recherche documentaire : la description du contenu intellectuel d'articles d'information.

[19-Lefèvre]
La recherche d'informations : du texte intégral au thésaurus / Philippe Lefèvre.- Paris : Hermès,2000.- 235 p.- ISBN 2-7462-0173-9
Les difficultés à trouver les informations, y compris sur l'Internet, sont surtout dues aux caractéristiques du langage. Partant de constat, ce livre fait l'inventaire des problèmes liés à la recherche d'informations ; puis il décrit les techniques appliquées pour les résoudre, des plus anciennes au plus récentes

[20-Collas et Chartron]
Logique conceptuelle et recherche d'information / Dominique Collas, Ghislaine Chartron.- Documentaliste-Sciences de l'information, janvier-février 1994, vol.31, n°1, p.9-15
Cet article démontre que l'efficacité d'une interrogation dépend très souvent d'un travail de conceptualisation qui doit être effectué avec la formulation de la requête : la logique conceptuelle consiste en l'explicitation systématique de tous les éléments notionnels, sémantiques et cognitifs contenus dans un sujet de recherche.

[21-Soergel]
Indexing and retrieval performance : the logical evidence / Dagobert Soergel.- Journal of ASIS, September 1994, vol.45, n°8, p.589-599
Cet article présente une analyse des caractères de l'indexation et de ses effets sur la performance de la recherche documentaire. L'auteur définit ce qui dans l'indexation affecte la recherche, à savoir : ses mécanismes, son exhaustivité, sa spécificité, son exactitude, sa cohérence. Il estime que la performance de la recherche dépend du rapport entre l'indexation et la question et de l'adaptation de la formulation de la question aux caractéristiques du système de recherche.

[22-Role]
De la lettre au sens : les recherches en texte intégral / François Role.- Documentaliste-Sciences de l'information, mai-juin 1993, vol.30, n°3, p.136-146
La recherche en texte intégral devient une réalité de plus en plus tangible : avant d'en présenter les principes, les techniques et les outils, cet article définit la notion de recherche en texte intégral par rapport aux autres types de recherche documentaire.

V/ Exemples concrets (8 références)

[23-Briand]
Usage des nouvelles technologies en centre de documentation de presse / Amélie Briand.- 2002.- 73 p. Mémoire DESS : Info-doc. : Paris, INTD/CNAM : 2002
Basé sur des enquêtes sur le terrain, ce mémoire répertorie, les modes d'organisation et de fonctionnement des centres de documentation de presse.

[24-Dalbin et Saleras]
Une expérience d'utilisation d'un système d'information documentaire en langage naturel / Sylvie Dalbin, Bruno Saleras.- Documentaliste-Sciences de l'information, décembre 2000, vol.37, n°5-6, p.312-324
Retour d'une expérience menée au Centre documentaire des Assurances générales de France (AGF), cette étude voudrait donner au lecteur les moyens d'évaluer les apports et les limites de la recherche en langage naturel, pour les usagers comme pour les documentalistes.

[25-Villacampa]
La création d'une base de donnés avec Retrievalware d'Excalibur / Alberto Villacampa. – 1999.- 43 p.- Mémoire DESS : Info-doc. : Paris, INTD/CNAM : 1999
Le présent mémoire a pour objet de définir des éléments de réponse à cette question : existe-t-il une opposition entre indexation manuelle et indexation automatique ? Il s'appuie sur une mission de stage à Air France qui a consisté en la création d'une base de données à l'aide du logiciel Retrievalware de la société Excalibur. Une méthode alternative est proposée qui combine indexation manuelle et automatique.

[26-Journées d'études de l'ADBS]
Outils linguistiques et nouvelles technologies : texte intégral des interventions [ressources électroniques] / Journées d'études de l'ADBS.- 19 décembre 1996 [consulté le 10 juin 2003].-
Nécessite Adobe Acrobat.- <http://www.adbs.fr/uploads/journees/619_fr.pdf>
Spécialement les interventions de Didier Rioux sur la complémentarité entre texte intégral et indexation documentaire : l'exemple du cédérom du Monde ; et celle de F. Bretonneau sur l'utilisation d'un langage normalisé dans la recherche en texte intégral, pourquoi, comment, exemple de la base de données du journal les Échos.

[27-Coret, Menon, Schibler et Terrasse]
Un système d'indexation structurée à l'INIST : bilan d'une étude préalable / Annie Coret, Bruno Menon, Danielle Schibler, Christophe Terrasse.- Documentaliste-Sciences de l'information, mai-juin 1994, vol.31, n°3, p.148-158
L'INIST envisage une approche linguistique intégrant un modèle d'indexation structurée : dans la phase d'analyse-indexation des documents, elle permet la suppression des problèmes de choix de vocabulaire, un gain de productivité, une amélioration de la qualité et de la cohérence ; dans la phase d'interrogation, elle permet de soumettre des requêtes en langage naturel et améliore la pertinence des réponses.

[28-Cotte]
Stratégie documentaire dans la presse / Dominique Cotte.- Paris : ESF éditeur, 1991.-(collection Systèmes d'information et nouvelles technologies.-122 p.- ISSN 0 298-3524
Spécialement le chapitre sur la fonction documentaire en presse (page 32 et suivante) qui définit bien les spécificités des services de documentation et les particularités de l'information et du discours de presse.

[29-Eyraud]
Évolution d'un langage documentaire de presse en vue d'une solution d'archivage électronique / Elizabeth Eyraud.- 1991.- 96 p. Mémoire DESS : Info-doc. : Paris, INTD/CNAM : 1991
Présentation du centre de documentation de l'Express. Étude de l'utilisation du plan de classement par la section "politique intérieure", comparaison avec le thésaurus utilisé pour la BDD interne et évolution vers l'archivage électronique.

VI/ Web sémantique (3 références)

[30-Archimag]
Accès à l'information : du web antique au web sémantique / dossier.- Archimag, juin 2003, n°165, p.21-29
Enrichir l'information qui circule sur le web et permettre à la machine de fournir à l'internaute des résultats de recherche plus pertinents que ceux qu'il obtient couramment : le web sémantique devrait aller bien plus loin que le web actuel. Le concept donne lieu à de nombreuses réflexions avec un accent particulier sur la question des langages et des ontologies.

[31-Rasmussen]
Language and representation : 3)Indexing and Retrieval for the Web / Edie M. Rasmussen.- Annual review of information science and technology, 2003, vol.37, p.91-124
Cette revue de la littérature explore les études en cours et celles des 5 dernières années sur l'indexation et la recherche documentaire sur le Web. Elle met l'accent sur les études qui développent et évaluent les techniques d'indexation et de récupération automatique de textes.

[32-Adams]
The Semantic Web : Differentiating Between Taxonomies and Ontologies / Katherine Adams.- Online, July-August 2002, vol.26, n°4, p.20-23
Le caractère « sémantique » du web : les « outils » hiérarchisés, les taxonomies et les ontologies.

ANNEXES

ANNEXE 1 : LES PRINCIPALES AMBIGUITES DU LANGAGE NATUREL[8]

Caractéristiques du langage naturel	Les difficultés dans la recherche d'informations	Définitions	Exemples
L'implicite	La pragmatique : Impossible à prendre en compte par des logiciels ou des langages documentaires	Liée au contexte du message, aux connaissances sur le monde, à l'usage… Domaine de la pragmatique : étude du « langage en action »	« Paul donna le billet à la jeune femme » Billet de banque ? Billet d'entrée ? Billet doux ?
La redondance	La synonymie	Mots ou expressions différents ayant le même sens ou des sens voisins	Voiture et automobile Tremblement de terre et séisme Train et chemin de fer
	La paraphrase	Expressions équivalentes mais de structure ou de termes différents	« Mon fils a cessé de fumer » « Jean a renoncé au tabac »
	Le glissement de sens	La dénotation : sens propre d'un mot La connotation : sens d'un mot dans un contexte particulier	« Il prend un bain » « Il est dans le bain »
L'ambiguïté	L'homonymie	Mots ayant la même forme, mais des sens différents	« Je porte la porte » « Les poules du couvent couvent »
	La polysémie	Mots ou expressions ayant plusieurs sens	Mémoire humaine Mémoire d'ordinateur Le mémoire de maîtrise
	L'homotaxie : Problèmes pour les logiciels de TALN	Une même syntaxe recouvrant des réalités différentes	« Jean est facile à convaincre » « Jean est habile à convaincre »

[8] http://www.uhb.fr/urfist/Supports/Rechinfo2/RechInfo2TALN.htm#Retour%20sur%20l'indexation

ANNEXE 2 : LE LANGAGE DOCUMENTAIRE CHEZ BAYARD

Une liste matière

1724 mots avec des notes d'application qui précisent le contenu et des « voir aussi » pour mettre en relation des termes proches

C'est une liste alphabétique de descripteurs

De nombreux descripteurs ont subdivisions particulières : sida (sida/pistes, sida/conférence, sida/France, sida/pays, sida/législation, sida/prévention, etc.) ; religion ; retraite ; cancer ; tennis ; voile ; etc.

Des listes « mono » ont été créées afin d'affiner un certain nombre de thèmes

- Listes fermées : aide à domicile (9 descripteurs), animal (201), art/manifestation (24), école supérieure (102), football (17), impôt (16), jeu (24), jeu d'argent (6), langues étrangères langues mortes (13), maladie (171), métier (231), ministère (31), musique (22), retraite régime (31), sport (88), tribunal (13), troisième age/hébergement (10)
- Listes ouvertes à création contrôlée : avion (51 descripteurs), médicament (54), religieux (69), secte (30), mouvements et partis politiques (création libre), associations et organismes (création libre)

Une liste de renvois

2632 mots, plus une liste de renvois pour l'Union Européenne et les listes « mono ».

Une liste d'entreprise

Toute organisation économique constituée pour la production de biens destinés à la vente, au négoce ou à la fourniture de services. Elles peuvent revêtir une forme financière, commerciale ou industrielle. On y trouve aussi les holdings, conglomérats ou groupes

Deux listes géographiques

Plus de 7000 descripteurs avec là aussi des listes « mono » ouvertes ou fermées

Géographie générale : comprend tous les pays et les zones géographiques (Asie du sud-est)

- Listes « mono » pour les îles, les fleuves, mers et montagnes et les organisations et conférences internationales (ouverte)
- Hiérarchie de 20 sous thèmes pour chaque pays et zones : pistes, généralités, catastrophe, culture, défense, économie, enseignement, environnement, histoire, immigrés, industrie, justice-criminalité, médias, politique étrangère, politique intérieure, politique sociale, religion, sciences, société, sports, terrorisme, transport, travail, villes et régions. Chacun des thèmes est défini (voir page 3)
- Subdivisions particulières pour certains pays : famille royale, histoire ancienne pour l'Égypte, etc.
- Pays en guerre avec liste particulière : conflit Israélo-palestinien, conflit État-Unis/Irak avec des subdivisions
- Union européenne avec subdivisions particulières et liste de renvois
- Québec, Tibet, Irlande du Nord, Palestine, Hongkong sont considérés comme des pays

Géographie France : villes, départements, régions, fleuves, îles et fleuves français

- Iles et fleuves/mono/
- Les régions, départements et villes sont subdivisés : pistes, généralités, culture, économie, politique, religion, société, sports, urbanisme
- Paris, Corse et Ile de France avec liste particulière

<u>Une liste biographie</u>
Particularités Dracula, Dieu, Jésus, troupe de théâtre, groupe de musique, personnage BD
Sous catégories : généralités, bio, déclarations, déplacements, sondages et des sous catégories plus importantes pour certaines personnes comme le Pape, Chirac

<u>Une liste à part pour les médias</u>
Plusieurs listes « mono » ouvertes

- Audiovisuel
- Bayard : Bayard/titres (77 descripteurs)
- Presse : presse titre/mono
- Radio : radio/émission/mono
- Télévision : télévision/émission/mono ; télévision/chaîne/mono ; télévision/satellite/mono

Contenu des subdivisions des pays

Pistes : Des informations synthétiques, des infographies, les repères de La Croix, cartes, données statistiques, chronologie

Généralités : Informations générales, plusieurs thèmes en même temps, une vue d'ensemble sur le pays, hors série, numéro spéciaux

Catastrophe : de toute nature, écologie, nucléaire, incendie. Trace en matière et en région et villes le premier jour ; pays/environnement peut être utile

Culture : art, cinéma, littérature, patrimoine, archéologie, tradition culturelle et manifestation culturelle à l'étranger

Défense : armée, défense, nucléaire, militaire, service secret, vente d'armes (guerre civile dans pays/politique intérieure)

Économie : agriculture, assurance, banque, bourse, commerce, conjoncture, finances, endettement, entreprise, monnaie, politique économique, tourisme

Enseignement : le corps enseignant, les écoles, étudiants, les études

Environnement : barrage, déchet nucléaire, écologie, faune, flore, patrimoine nature, protection de la nature

Histoire : anniversaire et commémoration

Immigrés : immigrés et pas émigrés

Industrie : industrie des biens de consommation et des biens de production, industrie extractive et de transformation, industrie lourde et légère (aéronautique, agroalimentaire, automobile, bâtiment, luxe, électronique, énergie, nucléaire, informatique, pétrole, sidérurgie, télécommunications, textile)

Justice Criminalité : système judiciaire, prison, police, droits de l'homme, criminalité, faits divers

Médias : audiovisuel, information, presse, publicité, radio, télévision

Politique étrangère : relations bilatérales politiques et économiques, ordre alphabétique (sauf France et Union européenne)

Politique intérieure : élection, gouvernement, hommes politiques, institutions, législation, parti politique, scandale et guerre civile

Politique sociale : système de santé, protection sociale, politique familiale

Religion : affrontements, enseignement, intégrisme, relation église-état, secte

Sciences : tous les domaines, médecine, géologie, recherche, etc.

Société : phénomènes de société (racisme, mode, loisir, drogue, etc.), la population (jeunes, retraités, sauf les immigrés)

Sports : les clubs ne s'indexent pas à la ville comme les clubs français

Terrorisme : actes terroristes et lutte contre le terrorisme (premier jour doublon en matière, particularités basque espagnol et français en matière)

Transport : routier, ferroviaire, etc.

Travail : monde du travail, condition de travail, emploi, salaire, chômage, syndicat

Villes et régions : un dossier est ouvert pour chaque région et chaque ville

ANNEXE 3 : COMPTE RENDU DE LA VISITE AU JOURNAL LE MONDE : RENCONTRE AVEC LE RESPONSABLE DU CENTRE DE DOCUMENTATION ET PRESENTATION DE LA BASE DE DONNEES (XYLEME ET SINEQUA)

1 - Contexte et historique

Le thésaurus du Monde a été créé en 1986 en interne au centre de documentation. Ce thésaurus est différent de la liste de classement des dossiers. Il comprend 3000 mots et 30 champs sémantiques avec 4 à 5 niveaux de hiérarchie. Les documentalistes étant spécialisées sur un domaine, chacune d'entre elles a voulu « garder ses mots ». De ce fait le thésaurus du monde est très précis mais aussi extrêmement lourd.

En 1987, une première informatisation est effectuée avec le logiciel Basis. Une base de données est créée avec l'intégralité du journal Le Monde, sans les images. Faite à partir du thésaurus, l'indexation sur cette base était très précise : 30 champs à renseigner (géographie, thématique, importance de l'information, etc.). A cette époque le centre de documentation est composé d'une équipe de 16 personnes. La moitié du temps de travail est consacré à l'indexation.

Depuis 2001, une réflexion sur une nouvelle informatisation est en cours. Au point de départ de cette réflexion, on trouve des critères économiques : 4 personnes partant à la retraite, le journal ne pouvait remplacer qu'un seul de ces postes. Il y a donc une obligation économique de gagner du temps à l'indexation. Cette automatisation est envisageable grâce à l'évolution des technologies et des moteurs de recherche, de même que l'intégration des images.

La base de données contient plus d'un million de documents (l'intégralité du journal Le monde et de ses suppléments depuis 1987). Le Monde avant 1987 est en cours de numérisation et pour cela la collection papier est sacrifiée (PDF simple sans recherche).

Dépouillement papier et dossiers : Le centre de documentation continue à dépouiller Le Monde sous format papier mais avec plus de six mois de retard. Les dossiers documentaires continuent à être alimentés par le dépouillement de près de 60 titres par semaine (hebdomadaires et mensuels compris). La sélection effectuée est très importante ; la liste des dossiers n'est pas réactualisée (une liste thématique et une liste géographique) ; chaque dossier regroupe le Monde et les autres sources. Il s'agit donc de gérer le passage des dossiers en informatique, ce qui n'est pas encore fait.

Les journalistes du Monde ont surtout besoin du Monde comme référence, les autres journaux subissent donc un dépouillement très sélectif. Les journalistes consultent surtout les dossiers biographiques (70% des dossiers empruntés) ; la documentation gère peu de questions ; les journalistes utilisent l'Intranet avec une interface simple (les brèves des deux derniers mois ; la dernière semaine du Monde visualisée en PDF ; les archives du journal depuis 1987, base de données avec Vérity, une interrogation simple sur le titre, la signature et le texte intégral).

Le produit cédérom est encore différent : il y a enrichissement à la base, le produit est fabriqué au Canada.
Enfin les archives du Monde accessibles sur Internet sont encore différentes (pas la même fabrication).

2 - Caractéristique et fonctionnement du système

Depuis le 24 décembre 2002, la nouvelle base de données du Monde est en test à la documentation (TIM : texte et image mode).
Un produit sur mesure a été créé issu de la coopération entre deux grandes sociétés : Xylème pour son savoir-faire en matière de base de données et notamment du format XML ; Sinéqua pour son moteur de recherche linguistique Intuition et pour son expérience dans la presse (Ouest France : avec Darwin de Cora à l'époque).
La version 1 en service depuis six mois à la documentation n'est pas accessible aux journalistes. La version 2 sera en test en septembre : les documentalistes ont demandé de nombreux changements à Sinéqua.
La mise au point de cet outil se fait en collaboration constante avec les deux prestataires.

- Xylème et la structuration XML
  Les qualités de Xylème sont principalement dans la très bonne qualité de sa base de données en XML qui permet de récupérer des documents déjà structurés.
- Les performances d'Intuition : traitements linguistiques et classification
  C'est un moteur d'indexation et de recherche qui permet une conceptualisation du langage grâce à l'utilisation à la fois d'algorithme mathématique et de traitements linguistiques.
  L'ensemble des documents est indexé automatiquement par le moteur après « le filtrage » des traitements linguistiques : gestion de la synonymie et de l'ambiguïté.
  Les dictionnaires intégrés sont la grande force d'Intuition : dictionnaire de la langue française et dictionnaire des synonymes avec une mise à jour tous les six mois en collaboration avec le responsable du centre de documentation qui peut demander l'intégration de certains synonymes (exemple altermondialistes synonyme d'antimondialistes). Les dictionnaires permettent une lemmatisation (ramener un mot à sa forme de base) et une analyse grammaticale qui facilite la désambiguïsation du langage naturel (analyse sémantique).
  Intuition peut également effectuer une phonétisation (recherche floue) qui permet de gérer les fautes d'orthographe à la saisie des requêtes.
  Intuition reconnaît et gère les mots composés et les expressions idiomatiques.
  La pertinence d'Intuition est basée sur des algorithmes mathématiques qui « jouent » sur la proximité, la fréquence, etc. Le tri par pertinence peut être paramétré à la recherche (40%, 60%, 80%) ou enlevé complètement (aucune). Le choix le plus fréquent étant 60%.
  La classification du lot résultat : des concepts sont extraits automatiquement et organisés par rapport à leur fréquence.

Une navigation dans ces concepts est possible par lien hypertextuel, ce qui peut orienter la recherche (croiser la requête avec un des concepts).

- **L'intégration des images**

  L'intégration des images est une des évolutions les plus importantes par rapport à l'ancien système. A chacun des articles est rattaché l'ensemble des images : photographie et infographie (carte, tableau). Ces images sont disponibles en JPG, et le module d'interrogation permet la recherche de ces images.

  Une autre image accompagne chacun des articles : un format PDF qui donne la vraie mise en page de l'article dans le journal. Il est très important de préserver cette mise en scène de l'article dans la page. « La mise en scène de l'information est également porteuse de sens. »

3 - Marquage ou indexation

200 articles par jour entrent dans la base de données ce qui correspond au Monde et à ses suppléments.

La structuration en XML de la base de données permet de « récupérer automatiquement » le titre, chapeau, auteur, place dans le journal et longueur de l'article. Néanmoins il y a encore nettoyage à l'entrée dans la base de données.

L'ensemble des articles du journal est marqué à l'entrée dans la base de données. Le responsable du centre de documentation parle de marquage et pas d'indexation, sans doute dans la mesure ou ce marquage est relativement sommaire ; en tout cas sans aucune mesure avec les 30 champs d'indexation de l'ancienne base de données. Par rapport à l'ancien système les champs personnes, thème/France, thème/pays ont disparus.

Il y a seulement 5 champs obligatoires et deux facultatifs :

1/Catégories : une liste de 50 catégories qui permettent de faire une typologie de forme des articles (éditorial, brève, etc.) et une typologie de contenu (bio, entretien, portrait, opinion, nécrologie, enquête, déclarations, etc.). Cette liste a été enrichie par l'ancienne liste des mots outils (chiffres, chronologie, historique, clés, évènements, etc.) et intègre une notion déjà utilisée auparavant (importance des articles). Ces catégories se retrouvent à la recherche et permettent d'affiner le questionnement.

2/Titres complémentaires : le titre de l'article est enrichi par les documentalistes en langage naturel et de façon complètement libre. Quelques lignes seulement pour lever l'ambiguïté des titres de presse et permettre une recherche sur la titraille (titre + sous titre + titre enrichi). Le titre complémentaire est jugé encore un peu long parfois par le responsable du centre. Les éléments les plus souvent rajoutés sont les noms de pays et des gens importants.

3/Genre œuvre : la notion d'œuvre correspond aux critiques. Dans le genre de l'œuvre on trouve cinéma, littérature, théâtre, etc.

4/Titre œuvre : le titre exact de l'œuvre

5/Auteur œuvre : le nom et le prénom de l'auteur.

6/Lien (facultatif) : pour faire le lien avec un rectificatif, un droit de réponse, une correspondance.

7/Commentaire (facultatif) : commentaire technique (absence d'images).

Les deux champs les plus importants sont les catégories et le titre complémentaire.
C'est un véritable changement des habitudes intellectuelles de l'indexation.
Pour marquer l'ensemble du journal 6 heures en moyenne (deux personnes en demi-journée ou une seule sur une journée).
Les champs « Œuvre » marquent une des spécialités du journal : identité.

Le dépouillement papier est très en retard pour Le Monde et extrêmement sélectif pour le reste de la PQN[9]. Les fichiers informatiques issus de la coopération documentaire ne sont pas encore exploités. Quand la PQN sera intégrée au système se sera sans indexation, seulement en texte intégral avec une localisation dans un dossier virtuel équivalent aux mots clés des dossiers papiers (une liste thématique et géographique avec dossier pistes).

4 - Les stratégies de recherche (voir graphiques p°6 et 7)

Le langage d'interrogation est assez simple avec + - * et «
Si Intuition gère sans problème les pluriels et les singuliers, les accents et les majuscules sont importantes dans la mesure où le traitement linguistique a besoin de tout ce qui véhicule du sens : mettre une majuscule à Noir lui permet de ne pas prendre en compte la couleur noire, de même pour les sigles et les entreprises.

La recherche sur l'article est en fait une recherche en texte intégral qui doit donc être affinée et précisée.
Il faut construire une stratégie en fonction des possibilités de l'outil. Il faut s'adapter, changer sa façon de voir et de chercher.
On peut affiner sa recherche de multiples façons. Certaines sont très classiques : recherche par le signataire, en fonction de la taille de l'article, par date.

Recherche en fonction du marquage des articles

- L'affinage par les catégories qui permet par exemple de rechercher un portrait, un entretien, etc.
- L'affinage par l'appartenance à un secteur ou à une rubrique du journal qui permet de chercher dans la rubrique Entreprise ou dans le secteur Économique.
- La recherche sur la titraille qui grâce au titre complémentaire est beaucoup plus pertinente.

Recherche en fonction des possibilités de l'outil

- La recherche avec images qui permet de trouver tableaux, cartes, etc.
- Le tri par pertinence qui permet de diminuer le bruit.
- La navigation dans les concepts permet d'affiner la recherche ; de « feuilleter »le lot résultat ; de trouver des idées ; d'autant plus que les concepts afficher comprennent ceux qui sont le moins présents dans le lot résultat et donc permettent de « voir les signaux faibles ».

[9] PQN : presse quotidienne nationale

Le maintien de l'historique : lien avec le thésaurus

- Toutes les notices de l'ancienne base entre 1987 et 2002 ont été récupérées et le thésaurus a été intégré. Il y a donc une possibilité de chercher avec les anciens mots clefs du thésaurus sur les notices d'avant décembre 2002.
- Apparemment cette fonction est très peu utilisée.

5 - Les plus et les moins

Les moins

- Le mélange des deux technologies a fait perdre à Intuition de sa finesse.
- C'est un système en texte intégral qui génère du bruit.
- Pas d'historique de recherche.
- Pas de troncature ni de sauf dans la titraille.
- Problème de traitement sémantique : pas de lien entre tauromachie et corrida.
- Le nombre de réponses ne s'affiche pas au-dessus de 250.
- La phonétique ne fonctionne pas toujours si les noms sont trop différents.
- Les traitements linguistiques ne traitent pas toutes les ambiguïtés : bavures policières.
- Ne pas pouvoir enlever un concept non pertinent, obligation de passer par la recherche et de faire un sauf dans le texte intégral.
- Ergonomie de l'interface : les pop-up.

Les plus

- Un produit sur mesure.
- Le gain de temps en passant de l'indexation au marquage.
- Le gain de pertinence que donne le marquage : indispensable.
- L'importance des dictionnaires.
- L'image que ce soit les infographies ou la vision de l'article en PDF.
- « Beaucoup de bruit c'est vrai mais en même temps on va plus vite à l'essentiel, on fait des dossiers plus pointus ».
- Rapidité malgré les traitements linguistiques qui prennent du temps.
- L'existence des concepts : ils ne sont pas tous pertinents, mais permettent au moins de gérer le bruit différemment (navigation, affinage, boite à idées, feuilletage et signaux faibles).
- Le tri par pertinence qui permet de diminuer le bruit.

Une constatation, cela implique une évolution du métier : « les documentalistes à terme ne s'occuperont plus des recherches simples (les journalistes les feront), elles alimenteront la base de données et prendront en charge les recherches plus complexes et auront peut-être plus de temps pour les dossiers ».

PRESENTATION DU LOT REPONSE SUR LA BASE DE DONNEES DU JOURNAL LE MONDE

NOMBRE DE REPONSES : 000

	PERTINENCE	DATE	AUTEUR	TITRE DES ARTICLES Avec plus ou moins chapeau et sous titre Avec plus ou moins titre enrichi	LONGUEUR	XML	PDF
CONCEPTS • • • ⊕							
CATEGORIES • • • ⊕							
SIGNATAIRES • • • ⊕							

INTERFACE DE RECHERCHE SUR LA BASE DE DONNEES DU JOURNAL Le MONDE

PARAMETRE DE RECHERCHE

- PERTINENCE : aucune / 40% / 60% / 80%

MODE D'AFFICHAGE

- NOMBRE DE RESULTATS : 10 / 20 / 50 / …
- TRI DES RESULTATS : date croissante / date décroissante / pertinence.
- TYPE D'AFFICHAGE : articles / œuvres

RECHERCHE GENERALE (taper sa requête)

- TITRAILLE : ☐
- ARTICLE : ☐
- SIGNATAIRE : ☐
- IMAGE ☒

DATE DE LA RECHERCHE

- DU 00 / 00 / 0000
- AU 00 / 00 / 0000

AUTRES CRITERES

- CATEGORIES : marquage 50 termes en pop-up
- TAILLE : brèves / court / moyen / long
- RUBRIQUES : les rubriques du journal en pop-up
- SECTEUR : les secteurs du journal en pop-up

ŒUVRES

- TITRE
- AUTEUR
- OEUVRE

HISTORIQUE

- MOTS CLES DU THESAURUS :

  Permet un lien avec les dossiers papiers et une interrogation avec les anciens mots clés…

ANNEXE 4 : COMPTE RENDU DE LA VISITE AU JOURNAL LES ECHOS : RENCONTRE AVEC LA RESPONSABLE DU CENTRE DE DOCUMENTATION ET PRESENTATION DES BASES DE DONNEES (BASIS+) ET DE LA COOPERATION DOCUMENTAIRE

Le centre de documentation des Échos travaille pour les 150 journalistes du journal. Trois grandes taches : gérer les archives du journal, constituer des dossiers documentaires et répondre aux questions des journalistes.
Les dossiers documentaires sont sous format papier, ils sont alimentés avec la presse nationale et les informations sur les entreprises recueillies sur le web notamment. Les dossiers sont empruntés par les journalistes.
Les dossiers entreprises constituent la part la plus importante des dossiers documentaires, ils sont fondamentaux pour les journalistes. Les dossiers ont une antériorité de trois ans.
Chaque été, les dossiers sont élagués par rapport aux dates et à l'actualité.

Le langage documentaire des Échos est un plan de classement alpha numérique qui est utilisé pour l'indexation papier de la presse nationale. Il comprend une vingtaine de grandes rubriques de la politique intérieure à la culture avec une spécificité économique.
Le plan de classement physique est en deux grands ensembles : d'une part les entreprises classées par nom qui correspondent au fonds le plus important pour le journal (spécialisation économique) ; d'autre part les dossiers généraux avec culture, politique, les personnalités, les informations internationales, etc.

La politique du centre de documentation depuis près de 4 ans est de réduire son fonds papier au maximum. D'abord en informatisant les archives des Échos, qui constituaient le plus gros du dépouillement papier. Ensuite par l'exploitation de la coopération documentaire entre les journaux de la presse quotidienne nationale.

La coopération documentaire entre les journaux de la presse quotidienne nationale est en fait un accord pour échanger gratuitement les fichiers électroniques des journaux. Cet accord, signé entre Libération, Le Monde, Le Figaro, La Croix, Les Échos et La Tribune, est effectif depuis un an.
Les fichiers sont envoyés pendant la nuit (sauf le Figaro, envoyés le soir après indexation), en format ASCII balisé et proviennent directement du système éditorial de chacun des journaux concernés.
Ce format permet de récupérer des articles déjà structurés avec le nom de l'auteur, la date, la place dans le journal, le titre, la taille.
Malgré les efforts d'harmonisation, un travail de nettoyage est obligatoire après la récupération des fichiers. En effet certains problèmes de compatibilité entre les systèmes obligent les documentalistes à vérifier l'exactitude des dates, des numéros de pages, des auteurs et la conformité du titre. Ce travail est effectué à partir des formats papiers des différents journaux.
Un journal en format électronique comprend environ 150 à 200 fichiers (138 fichiers pour Le Figaro).

Ce système est encore en test aujourd'hui. La récupération de certains des fichiers ne fonctionne pas encore, comme pour Le Monde notamment. Les nombreux changements de système éditorial posent encore des problèmes.
A terme, chacun espère que les accords seront étendus à la presse hebdomadaire. Dans cette perspective une troisième base de données est envisagée.

Le dépouillement papier du journal Les Échos n'est plus effectué depuis près d'un an.
Par sécurité certains articles très importants sont encore conservés en papier dans les dossiers documentaires : synthèse sur une entreprise, grandes biographies.
De même, les infographies sont encore conservées même si aujourd'hui leur récupération automatique est au point. Les difficultés ont persisté longtemps sur ce point, et on s'assure de ne rien perdre de cette richesse du journal.
Le reste de la presse quotidienne continue d'être dépouillé manuellement dans la mesure où le système de coopération documentaire n'est pas encore complètement au point.
L'indexation se fait à partir du langage documentaire ; c'est une indexation légère qui comprend un ou deux mots clés au maximum. Les documentalistes étant spécialisés sur un domaine, chacun dépouille en fonction de ce domaine.

Il existe deux bases de données séparées et gérées par le même outil (Basis+), une pour les archives du journal et l'autre pour la coopération documentaire.

La base de données des archives des Échos comprend les archives du journal depuis 1991. Les fichiers sont en format HTML et sont disponibles sur l'Intranet par l'intermédiaire de Web Basis.
Les journalistes utilisent énormément cette base pour leurs recherches. La majorité de ces recherches consistent à retrouver un article des Échos, qu'ils en soient l'auteur ou non. Ils effectuent ce genre de recherche seul, ce qui permet aux documentalistes de dégager du temps pour les recherches plus complexes.
Le moteur de recherche intégré permet les requêtes booléennes classiques sur le titre ou le texte intégral, le croisement des requêtes, la recherche par auteur, date, rubriques du journal, typologie d'articles, etc.
Si les journalistes ont eu du mal à s'habituer à l'absence des Échos dans les dossiers documentaires papiers, ils sont apparemment aujourd'hui satisfaits d'utiliser la base de données.
Tous les articles du journal sont conservés ainsi que les infographies qui sont très importantes aux Échos (tableaux statistiques, courbes financières, etc.)
L'indexation de la base est faite manuellement et correspond à 70% au langage documentaire du centre.
De plus en plus il y a un allégement de l'indexation sur la base des échos, avec un souci de faire évoluer les index vers le plus précis. Les infographies sont indexées par un documentaliste. La relecture de la base se fait en 30 minutes environ.

La base de la PQN[10], est séparée de celle des archives du journal. Elle est gérée par le même logiciel Basis+. C'est une base en format de fichiers HTML, dont l'ergonomie est très simple.

Une sélection est effectuée sur l'ensemble des fichiers de la PQN mais cette sélection est beaucoup moins importante que celle effectuée pour le dépouillement papier.
Tous les articles économiques sont conservés ainsi que les grands papiers en politique intérieure, internationale et en culture. Ne sont pas retenus : les brèves, le courrier des lecteurs, la télévision, le sport. Les articles non sélectionnés sont « écrasés » définitivement, ils n'apparaîtront plus dans la base de données.
Aucune indexation n'est effectuée pour l'instant sur la PQN. Après vérification de la « propreté » des articles, les documentalistes assurent un enrichissement très rapide du titre. Pour les brèves, les 20 premiers mots du texte apparaissent automatiquement dans le cadre du titre, et le documentaliste choisi de garder ou d'enrichir ce choix. Les infographies ne sont pas traitées pour l'instant.
L'objectif à terme est de constituer des dossiers virtuels. Une indexation très simple sera effectuée, deux mots-clés au maximum, pour créer des dossiers en ligne qui auront un lien avec les archives des Échos. Ce projet est encore au stade de la réflexion.
Cette base de PQN n'est pas interrogeable par les journalistes. Ce sont les documentalistes qui font des recherches pour eux et qui impriment les articles si besoin. Les recherches sur la base se font uniquement en texte intégral. Les dossiers documentaires papiers sont encore beaucoup utilisés.

L'organisation du travail a changé depuis un an en fonction des nouvelles données.
Les quatre documentalistes se sont spécialisés sur des domaines précis : un travail sur les services et l'international ; un sur le high-tech et la macro-économie ; un sur la banque, les finances et l'industrie ; un sur la culture, la politique et les « idées ».
Pour compléter l'équipe, on compte deux aides documentalistes, un documentaliste spécialisé sur le web et un documentaliste à mi-temps qui travaille sur Les Enjeux. Deux étudiants viennent le soir afin d'assurer un service de « push » sur l'intranet.
Le temps consacré à l'indexation a diminué : chaque documentaliste s'occupe de son domaine pour ce qui est du dépouillement papier et traite un titre de la PQN sur la base. Les Échos n'étant quasiment plus traités en papier, le travail d'indexation est normalement terminé à la mi-journée.
Avec l'explosion de l'Internet cela a contraint les gens à travailler autrement, à sélectionner plus pour se démarquer des recherches sur le web. Il ne s'agit plus aujourd'hui de donner aux journalistes des gros dossiers et les laisser chercher, mais de faire des recherches pointues à forte valeur ajoutée.
Comme les journalistes empruntent les dossiers documentaires, il faut aller au devant de leurs questions et maintenir une relation humaine qui est fondamentale dans le travail. La spécialisation de chaque documentaliste sur un domaine à permis de renforcer les liens avec les journalistes qui ont un interlocuteur privilégié..

[10] PQN : presse quotidienne nationale

ANNEXE 5 : COMPTE RENDU DE LA VISITE AU JOURNAL LE NOUVEL OBSERVATEUR : RENCONTRE AVEC LA RESPONSABLE DU CENTRE DE DOCUMENTATION ET PRESENTATION DE LA BASE DE DONNEES

Les services du Nouvel Observateur emploient 240 personnes au total dont 150 journalistes (pigiste compris).
La photothèque comprend 8 iconographes.

Présentation du service de documentation

Le service de documentation compte 10 rédacteurs-documentalistes et 1 archiviste.
Leur titre est celui de Rédacteur-documentaliste car tous écrivent des articles dans le journal. Par conséquent ils sont détenteurs de la carte de presse.
Leur travail quotidien se fait en collaboration avec les journalistes du journal : dépouillement de la presse; constitution de dossiers thématiques, réponses aux demandes des journalistes, rédaction d'articles ou de brèves, présence aux conférences de rédaction.

Chaque rédacteur-documentaliste travaille par grands secteurs : Politique Étrangère (regroupe également les médias, la justice, les droits de l'homme et la culture) ; Politique Intérieure (1 documentaliste gère les personnalités françaises, une autre les syndicats, les transports, l'énergie, les banlieues, emploi/chômage) ; Notre époque (faits de société) ; Europe ; Économie (2 documentalistes gèrent la macro-éco et la micro-éco) ; Social ; Culture (arts et spectacles).

Le fonds documentaire

Essentiellement un fonds papier composé de dossiers thématiques élaborés par les documentalistes à partir du dépouillement et de l'indexation de la presse écrite : 15 journaux de la presse quotidienne française et étrangère ; 20 hebdomadaires français et étrangers ; 10 mensuels français et étrangers ; une sélection de la presse quotidienne régionale et des revues spécialisées (les lettres d'information…).

Rajouter à cela une bibliothèque constituée de centaines de livres et d'usuels (Who's who ? , L'État du Monde, etc.) ; l'utilisation d'Internet et l'abonnement à des bases de données externes de presse comme l'Européenne de données et Europresse.

De la gestion des archives a l'Informatisation du Service de documentation

Les archives

Une seule personne gère les archives. L'archiviste travaille en étroite collaboration avec les documentalistes. Son rôle est d'indexer le journal et de répondre aux demandes extérieures. La collection comprend l'ensemble du Nouvel observateur depuis 1993 en format PDF. La numérisation des journaux plus anciens est en cours, avec la volonté de remonter jusqu'en 1964.
La direction du journal s'est orientée vers une politique de commercialisation des archives, voilà pourquoi on trouve sur le site du Nouvel Observateur le journal depuis 1993 indexé en texte intégral.

Informatisation du service de documentation

Pour pouvoir exploiter les fichiers de la coopération documentaire, il fallait informatiser le service de documentation.
Le choix s'est porté sur la société Eurocortex et sur un logiciel documentaire de gestion de contenu. Après plusieurs mois de test, la base de données gérée par le logiciel ICM (Intelligent Content Manager) fonctionne depuis mars 2003. Pour mettre en place ce logiciel, il a fallu au préalable de nombreuses réunions de travail entre la responsable du service de documentation et les informaticiens mobilisés, afin de définir ensemble une architecture et un accès à l'information. Avec cette application l'objectif visé est d'assurer le lien entre l'affluence de données hétérogènes circulant dans le journal et la diversité des supports de diffusion.

Le thésaurus

La perspective d'informatisation du service de documentation a nécessité une réflexion autour d'un thésaurus spécifique. En effet, jusqu'alors l'indexation était effectuée à l'aide d'une « Nomenclature par secteur ». Cette nomenclature correspondait au plan de classement physique des documents, c'est-à-dire aux dossiers thématiques. Ce langage très simple est encore aujourd'hui utilisé par l'archiviste pour indexer le Nouvel Observateur.
8 mois de travail qui a mobilisé l'ensemble des rédacteurs-documentalistes pour constituer cet outil d'indexation et de recherche en ligne. La plus grande difficulté a été de transposer ce thésaurus via un logiciel documentaire tout en conservant l'esprit et le mode de classification papier.
Le thésaurus comporte près de 1200 termes, répartis en champs sémantiques et hiérarchisés sur trois niveaux au maximum. Depuis la mise en service de la base de données, des termes ont déjà été rajoutés. La gestion et la mise à jour du thésaurus sont quotidiens. Un espace réservé aux candidats descripteurs a été créé dans la base de données.

La base de données

Les documentalistes sélectionnent et indexent l'ensemble des documents entrant dans la base, en fonction de leurs domaines de spécialisation. La sélection s'effectue sur les hebdomadaires appartenant à la coopération documentaire dont les articles sont récupérés automatiquement. Comme aux Échos, cette récupération nécessite une vérification et un nettoyage.
Les documentalistes font également une sélection des articles du Nouvel Observateur.

Pour les autres journaux, notamment les quotidiens, les documentalistes vont chercher sur les sites en ligne, les articles dont elles ont besoin. Cet arrangement est provisoire, en attendant la signature d'une coopération entre les hebdomadaires et les quotidiens. Le dépouillement papier est de plus en plus succinct, il concerne les publications quotidiennes et étrangères qui ne peuvent être récupérées en ligne.
A noter que la sélection des articles est très importante ; seuls sont conservés les articles d'une certaine importance, les synthèses, les portraits et interviews.
Les journalistes ont accès à la base de données par l'Intranet de l'entreprise.

L'indexation

La base de données cumule une structuration par champs et une indexation automatique du texte intégral. Les champs récupérés automatiquement sont : la source, la date, l'auteur, la page, la taille (en signes), le titre et le chapeau. Sont liés à l'article sans être indexés, les encadrés et les visuels.

L'indexation humaine est assez importante. Elle comprend 11 champs : rubrique, personnalité, entreprise, région du monde, pays, région, ville, mot outil, thème (3 champs).
Le thésaurus est entièrement intégré : à la fois les thèmes et les mots outils. A noter que dans la presse, les mots outils contiennent également une typologie des informations contenues dans l'article : chronologie, statistiques, déclarations, entretien, etc.
Le thème correspond au sujet de l'article. L'indexation est faite avec l'aide du thésaurus en ligne. En général deux ou trois termes sont sélectionnés en essayant d'aller vers le plus précis.

La recherche

Il existe deux interfaces de recherche : une simplifiée pour les journalistes, une recherche avancée pour les documentalistes. Les termes de la requête sont surlignés.
La recherche simplifiée permet aux journalistes de faire une requête sur le titre et chapeau, d'interroger sur l'auteur et la date et de poser une question sur le texte intégral. 30% des journalistes utilisent cette base, principalement pour des recherches simples comme retrouver un article.
La recherche avancée n'est accessible qu'aux documentalistes. Tous les champs de l'indexation sont paramétrables à la recherche.

NOTICE DU MÉMOIRE

Langages documentaires et nouvelles technologies : l'avenir des langages et leur positionnement au cœur des systèmes d'informations dans le contexte de la presse./ Odile Contat.- 2003.- 89p. Mémoire DESS : Info-doc. : Paris, INTD/CNAM : 2003

Résumé :

L'auteur définit dans un premier temps les fonctions du langage documentaire (normalisation, désambiguïsation, organisation des connaissances et structuration), et le fonctionnement de l'indexation automatique et des différents traitements linguistiques, statistiques et sémantiques.
Puis il replace la problématique de la recherche d'information dans le contexte particulier de la presse et décrit les systèmes d'informations de Bayard, du Monde, des Échos et du Nouvel Observateur.
Ensuite, en partant des besoins d'informations des journalistes, il tente de définir les rôles et les fonctions des langages documentaires face à l'utilisation des nouvelles technologies en fonction d'une typologie de l'indexation.
Dans les systèmes d'informations presse, les outils d'indexation automatique et de recherche en texte intégral peuvent prendre en charge le traitement linguistique et sémantique, la classification automatique, etc.
Pourtant le rôle des langages documentaires et la place de l'indexation humaine restent importants. Un usage sélectif de l'indexation manuelle permet une meilleure conceptualisation du sujet et une véritable structuration des documents en fonction des types d'information. L'intégration des langages améliore les performances de la recherche sémantique et permet une meilleure normalisation et désambiguïsation du langage naturel. Les fonctions d'organisation des connaissances, de médiation et d'aide à la recherche des langages documentaires viennent en complémentarité des classifications et de l'hypertexte.
Le langage documentaire apparaît en complémentarité de la recherche en texte intégral. Même s'il devient de plus en plus invisible, intégré aux nouvelles technologies, ces fonctions sont préservées et demeurent indispensables.

Mots clés[11] : langage documentaire, langage naturel, accès à l'information, système de recherche d'information, TAL (traitement automatique du langage), indexation en texte intégral, surindexation, presse

[11] L'indexation de ce mémoire a été faite à partir d'un thésaurus de l'information documentation; effectué par des élèves (Rabéa Chakir Trébosc, Floriane Gauffre, Odile Contat, Alina IvanciucDeniau) dans le cadre de l'INTD durant l'année 2002/2003 et à destination de l'ADBS.