Avec le développement des intranets, nombre de centres de documentation évoluent d'un modèle « classique », avec un logiciel de gestion et de recherche de références bibliographiques, vers un système d'information en langage naturel directement accessible aux usagers. Retour d'une expérience menée au Centre documentaire des 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. Elle met aussi l'accent sur le nécessaire « décentrage » – par rapport à nos pratiques et habitudes – qu'impose le développement de tels dispositifs orientés « usager final ».

par **SYLVIE DALBIN**
consultante, ATD, Paris
et **BRUNO SALLÉRAS**
responsable du Centre documentaire des AGF

Une expérience d'utilisation d'un système d'information documentaire en langage naturel

APRES AVOIR UTILISE PENDANT HUIT ANS un logiciel de gestion et de recherche de références bibliographiques, les Assurances générales de France (AGF) ont opté en 1998 pour un système documentaire en langage naturel supporté par le logiciel Spirit de la société T-GID[1].

Permettre à l'usager d'effectuer lui-même directement sa recherche d'information est une pratique ancienne aux AGF. Dès 1991, huit cents personnes (dont, en fait, deux cents usagers réguliers) étaient habilitées à interroger la base documentaire, Sydoc, sous Basis Plus. Mais les études menées auprès des personnes – utilisatrices ou non de Sydoc – qui fréquentent le Centre documentaire des AGF (voir page 320) et l'évaluation de l'utilisation de cette base ont montré que, souvent, un système fondé sur une recherche par mots contrôlés, opérateurs booléens et syntaxe rigoureuse est faiblement et mal employé par les non-professionnels de la documentation.

Tout d'abord le choix des termes leur est difficile, surtout lorsque ce choix n'apparaît pas dans l'index des termes proposé à l'écran ; le concept d'équivalence documentaire et les principes de normalisation des termes (pré ou postcoordination) sont en effet mal appréhendés par les non-spécialistes. Le passage du ou des termes choisis

par l'usager pour exprimer sa demande au terme pertinent du langage documentaire, choisi par le documentaliste, apparaît à chacun comme une étape délicate, et surtout très aléatoire : le non-expert du domaine sera surpris par une terminologie différente de la sienne, voire inconnue de lui ; l'expert regrettera le choix de termes de sens approchant mais non sémantiquement équivalents.

En second lieu la syntaxe de la requête, nécessairement rigoureuse, déconcerte l'usager. Les opérateurs booléens sont souvent inconnus, mystérieux ou encore mal utilisés, même aujourd'hui, à l'ère d'Internet et des moteurs de recherche. Or le principe même d'un dispositif s'appuyant sur un thésaurus est de poser des questions précises en effectuant des combinaisons de termes, l'usage d'un seul terme générant beaucoup de bruit.

La recherche en « langage naturel », en revanche, à la fois simple et assistée grâce aux traitements linguistiques et au principe de reformulation automatique des questions (voir page suivante), semblait parfaitement adaptée, comme mode d'accès à l'information, aux usagers de l'intranet documentaire des AGF, non-spécialistes de la recherche documentaire.

Les études portant sur le système d'information documentaire (SID) et son utilisation, études qui ont conduit au choix du langage naturel comme outil d'indexation et recherche, ont également permis de faire évoluer le concept même de SID. En effet, il est apparu que les difficultés d'utilisation du système documentaire « classique » n'étaient pas toutes imputables à la rigidité des langages de commande ni aux difficultés de choix des termes de recherche : la base documentaire représentait un modèle monolithique d'organisation et de représentation des documents, inadapté au traitement de certains types de documents ou d'informations, et inapproprié à certaines formes de questionnements et de besoins des usagers.

Pour résoudre cet aspect du problème de la recherche d'information, en particulier celui de l'orientation de l'usager vers la « bonne » ressource informationnelle, un important travail de reconception et de restructuration de l'intranet documentaire[2] a été mené par l'équipe des AGF, resituant les bases de documents susceptibles d'être interrogés en langage naturel dans un « espace informationnel »[3] élargi et mieux structuré.

Le présent article s'appuie sur l'étude des résultats de deux années (entre septembre 1998 et décembre 2000) d'utilisation des bases documentaires de l'intranet documentaire des AGF interrogeables en langage naturel sous Spirit. Il ne constitue pas une analyse des technologies linguistiques sous-jacentes, ni une évaluation du logiciel Spirit, ni une étude comparative des logiciels en texte intégral et en langage naturel. Nous avons souhaité montrer, par des exemples concrets, ce que recouvrent les démarches de recherche en langage naturel, ainsi que leurs avantages et leurs limites pour les usagers et pour les documentalistes.

Sylvie Dalbin est consultante au sein de la société Assistance & Techniques Documentaires, membre du GIE Desybel (34 boulevard des Italiens, F-75009 Paris, téléphone +33 (0)1 42 78 03 70, télécopie +33 (0)1 42 77 18 10, courriel SylvieATD@aol.com).

Bruno Salléras est responsable du Centre documentaire des Assurances générales de France (AGF, 87 rue de Richelieu, F-75002 Paris, téléphone +33 (0)1 44 86 20 10, télécopie +33 (0)1 47 03 93 11, courriel sallera@agf.fr).

Cette présentation est également l'occasion d'insister sur l'importance de repositionner la base documentaire, son contenu et les modalités d'accès à l'information, dans le contexte d'une offre globale de services, nécessairement multiple par sa forme et ses modalités d'usage. Ce qui suppose une remise en cause de ses pratiques professionnelles...

Dans une première partie, nous présenterons l'organisation des ressources documentaires sur l'intranet des AGF puis, dans une deuxième partie, les résultats de l'utilisation de cet intranet documentaire à partir de questions posées par les utilisateurs sur les bases documentaires.

1 L'intranet documentaire des AGF

Un espace informationnel réorganisé

Début 2001, environ quatorze mille documents sont mis à la disposition de l'ensemble des collaborateurs des AGF. Les bases documentaires accessibles par la rubrique *Ressources documentaires* sont composées d'articles de presse, ainsi que de références bibliographiques de monographies (ouvrages, études) accompagnées de sommaires. Elles se segmentent suivant deux axes : une base « généraliste » (*Intradoc*) sur les domaines relevant de l'entreprise, de l'économie et des finances, des ressources humaines, du marketing et du droit,

1 Un article récemment paru ici même explique les raisons de ce choix et présente l'intranet documentaire actuel : La nouvelle conception de l'intranet documentaire des AGF, *entretien de Sylvie Dalbin avec Bruno Salléras, septembre 2000, vol. 37, n° 3-4, p. 200-204.*

2 En ce qui concerne l'interaction entre le système et l'usager, voir également l'article de Joëlle Cohen paru dans le dernier numéro de cette revue : L'écran efficace : trois lois fondamentales de la perception visuelle, *septembre 2000, vol. 37, n° 3-4, p. 192-198.*

3 Le terme « espace informationnel » renvoie à l'offre de services et de produits proposés ici sur le site intranet, mais il suppose également, en tant qu'espace, qu'il existe une « signalétique » et un « mode d'emploi » de cette offre.

Indexation et recherche en langage naturel

Le vocabulaire utilisé depuis une quinzaine d'années pour caractériser certains systèmes de recherche documentaire a parfois été malmené. Aussi paraît-il important de préciser quelques termes utilisés dans cet article à propos de l'indexation et de la recherche dites « en langage naturel », ainsi que d'expliquer en quoi consistent les traitements automatiques qui sont mis en œuvre.

Dire que l'on propose aux usagers un accès à l'information en langage naturel signifie que :

- l'indexation et la recherche portent sur la partie textuelle de documents, et non uniquement sur des références de documents ;

- les utilisateurs formulent leur question dans leur langage, sans contrainte terminologique (termes choisis ou mots clés) ni syntaxique (formulation de la question) ;

- des traitements automatiques de nature linguistique et statistique appropriés enrichissent l'expression de la question (indexation de la requête) et assurent automatiquement l'étape d'appariement entre les textes et la question (recouvrement entre les deux index résultant du traitement des textes et de la question).

Les logiciels qui offrent des traitements linguistiques proposent, en plus, des possibilités de classification des résultats de la question. Dans Spirit, ce regroupement en « classes » correspond à une combinatoire (booléen pondéré) de toutes les questions qui pourraient être formulées, combinatoire optimisée grâce à un algorithme dédié. Cette organisation des résultats constitue une aide précieuse pour les usagers dans l'étape de sélection des informations, étape généralement occultée dans les systèmes classiques.

L'expression « recherche par le contenu » (*content retrieval*) élargit, quant à elle, la problématique de recherche des textes *(text retrieval)* à l'image, au son, au mouvement. On peut en effet vouloir retrouver une photographie par les objets qu'elle contient, une musique par un son particulier. D'autres techniques que celles utilisées pour la recherche dans des textes sont ici utilisées, mais le principe de base est le même : retrouver une information ou un document directement par son contenu.

Le récent engouement pour ces logiciels et la nouvelle dénomination commerciale* de certains d'entre eux pourraient laisser penser, à tort, que leur origine est récente. En réalité, certains font partie du paysage documentaire depuis vingt ans. Mais il est vrai que la gestion électronique des documents dans les années quatre-vingt-dix puis, à l'heure actuelle, les possibilités de recherche dans Internet ont donné un puissant élan à ces outils. Parallèlement, les éditeurs de logiciels ont dû réaliser certaines évolutions non pas tant fonctionnelles que de mise en œuvre (simplifiée) et de coût (compétitifs pour certains).

Traitements linguistiques et statistiques

Dans l'indexation en « langage naturel », contrairement à l'indexation en « texte intégral » qui ne fait qu'isoler des chaînes de caractères (unitermes entre deux blancs), différents niveaux de traitements automatiques de nature linguistique sont mis en œuvre.

• **Prétraitement linguistique.** Ce premier traitement permet de segmenter et formater les corpus de texte en phrases, de corriger les erreurs typographiques ou d'orthographe par exemple (et de leur associer des identifieurs qui seront utilisés dans les étapes suivantes).

• **Traitement morpholexical et morphosyntaxique.** Il permet de traiter les multiples variantes d'un même terme (singulier/pluriel ; adjectif/adverbe/verbe/substantif) par les fonctions de lemmatisation** et de structuration grammaticale, et d'identifier précisément le sens d'un mot grâce à son environnement dans la phrase (mots composés et expressions, désambiguïsation).

• **Sémantique générale de nature lexicale.** Ce niveau de traitement concerne le réseau de relations qui rapproche des concepts les uns des autres ; il permet de diminuer le silence et le bruit en traitant les familles de mots, la synonymie, l'hyperonymie (meuble/siège), la métonymie (partie de) ou l'association.

• **Sémantique contextuelle (ou pragmatique).** En traitant le sens au niveau des phrases et de leurs interrelations, cette analyse sémantique améliore le traitement du bruit et du silence à partir de règles linguistiques et de connaissances pragmatiques. On réduit ainsi l'ambiguïté de certaines phrases dont le sens dépend de leur contexte (comprendre l'implicite par exemple).

• **Traitement statistique.** Des algorithmes spécialisés permettent, à partir des données générées par les traitements statistiques, de pondérer les termes retenus dans les étapes précédentes.

• **Traitement de regroupement/classification.** Les résultats sont organisés par classes de documents en fonction de leur plus ou moins grande « pertinence » par rapport à la question posée, c'est-à-dire de leur degré d'appariement calculé.

NB. La version de Spirit utilisée possède un dictionnaire complet de la langue d'environ six cent mille entrées, auquel on peut adjoindre un dictionnaire « privé » (ici d'une centaine de termes). Dans l'application aux AGF, Spirit assure les trois premiers niveaux de traitement linguistique : lexical, syntaxique, sémantique générale de nature lexicale, complétés par des traitements statistiques et de regroupement.

---

* Intuition (anciennement Darwin, de Cora) ou Lexiquest (anciennement Erli).

** Lemme ou forme canonique : formes de base des mots, par opposition à la forme fléchie qui représente les différentes variantes d'un mot.

De l'énoncé de la question à l'exploitation des résultats : les différentes étapes

Étape 1 - Compréhension de l'architecture et de l'organisation du dispositif documentaire :
présélection (de la rubrique du site, du domaine, du type de documents, etc.)

- Site *Ressources documentaires* signalé sur l'intranet en trois endroits différents.
- Structuration de l'espace en fonction des pratiques utilisateurs
- Terminologie adaptée

Étape 2 - Capture de la requête
- Élaboration de l'équation initiale :
- dialogue en langage naturel
- recherche multicritères sur quelques champs structurés précis
- Aide au choix des termes :
- liste des thèmes, une sélection de dix titres de revues (les principales), des produits, des entreprises
- rétroaction de pertinence (*relevant feedback*) ou « recherche intuitive » : possibilité d'utiliser le texte d'un document comme requête

Étape 3 - Traitement de la requête (Spirit)
- Dictionnaire « privé » (propre à l'entreprise utilisatrice)
- Tolérance aux erreurs (de manipulation, orthographiques)
- Reformulation (automatique). Exemple : tondeuse à gazon = tondant sa pelouse

Étape 4 - Présentation des résultats
- Présentation du lot résultat : tri par classes (regroupement de documents en fonction du degré de pertinence)
- Présentation des informations résultats et circulation au sein des documents :
- distinction par « unité d'information » (document, partie de document) et navigation entre elles
- positionnement sur le meilleur regroupement de mots recherchés et son contexte (surbrillance)
- document joint en PDF

fortement orientée vers les problématiques de l'assurance ; et une base plus spécialisée (anciennement dénommée *Brèves*) sur les marchés de l'assurance, les clients et les produits (*Concurrence*).

L'accès à ces deux bases documentaires, qui constituent la partie originelle de la rubrique *Ressources documentaires*, a été repensé et réorganisé à l'occasion d'une réflexion plus large qui portait sur l'offre de services et produits documentaires, et sur la composition et l'organisation de la rubrique[4].

Les besoins exprimés par les utilisateurs au cours d'enquêtes menées depuis trois ans ont permis de caractériser des demandes assez différentes dans leur nature pour envisager une restructuration de l'espace informationnel et des accès distincts à ces différentes ressources.

Au niveau des domaines couverts, en particulier pour l'information juridique. Le rythme et les spécificités des demandes portant sur l'information juridique, les particularités des démarches de questionnement des utilisateurs dans ce domaine, le circuit de traitement des documents contrôlé par des documentalistes-juristes ont permis d'envisager une solution particulière pour ce type d'information[5] : un portail juridique (actualités parlementaires vues à travers la presse, veille juridique, contentieux du groupe AGF, codes en ligne, etc.). Ce portail, baptisé *Actualités juridiques* et accessible à partir du site *Ressources documentaires*, est apparu en décembre 2000[6].

Dans la même démarche de restructuration par domaines, la base documentaire de références bibliographiques d'origine, *Sydoc*, a été remaniée[7] de telle sorte que son contenu (articles et ouvrages) est aujourd'hui accessible par deux rubriques : *Poser une question* (en langage naturel) et *Interroger par thème* (dans un choix d'une trentaine de thèmes sélectionnés). Par ailleurs, une autre base, *Concurrence*, également issue de *Sydoc* et regroupant les articles (mais pas les ouvrages) qui traitent de compagnies d'assurances, de bancassurance, etc., est accessible par la rubrique *Surveiller les concurrents*. En effet, les demandes liées à la concurrence se traduisent par des questions portant soit sur une société, soit sur une activité transversale (stratégie, résultats, communication, politique sociale, produit). Dans la base *Concur-*

4 Cette rubrique a été essentiellement élaborée et construite par Lydie Pineau, du Centre documentaire des AGF, et Monique Giuly, d'AGF Informatique.
5 Les ouvrages sur le droit restent dans la base générale.
6 Cette rubrique a également été élaborée principalement par des documentalistes.
7 Seuls les ouvrages (avec résumés) ont basculé de Sydoc *vers la nouvelle base* Intradoc. *Il n'y a pas eu de reprise des références des articles. À partir de juin 1998, les articles sont intégralement numérisés ou téléchargés dans la base Intradoc.*

rence, il est apparu plus efficace (recherche de qualité et d'une solution plus simple et plus rapide pour les usagers) de proposer une sélection de produits et de sociétés plutôt qu'une recherche en langage naturel, qui reste néanmoins possible.

Au niveau de la nature de l'information fournie par le système. Un document textuel ne constitue pas toujours la réponse à un besoin documentaire. L'analyse des questions montre que la « nature » de l'information (une donnée ou un argumentaire, par exemple) constitue un critère de recherche, sous-jacent à la recherche même, pas toujours formulé explicitement. Dans la pratique professionnelle, des articles, souvent de durée de vie courte, sont insérés dans les bases documentaires parce qu'ils contiennent des faits, des données chiffrées, un tableau (indice de la consommation, classement de sociétés, par exemple). Les contenus de ces documents sont rarement représentés de façon efficace par les mots clés, ce qui rend leur recherche difficile.

Ce besoin d'un accès rapide et direct à des données chiffrées rend nécessaire la construction de bases d'information spécifiques. Celles-ci génèrent une activité documentaire d'exploitation de documents et d'extraction d'information pour l'alimentation de deux autres rubriques du site, *Indices* et *Classements* (des compagnies d'assurance), réalisées par les documentalistes. Ce travail permet de fournir immédiatement des réponses précises aux questions de cette nature, et épure d'autant les bases documentaires de documents qui pouvaient être considérés comme générateurs de bruit pour certaines questions.

Ainsi, à la place d'un mode d'accès unique à un fonds unique et homogène dans sa représentation (notices organisées en champs structurés, décrivant l'ensemble des documents que possède le Centre documentaire), nous proposons des accès distincts à des fonds distincts : des listes factuelles d'indices socio-économiques et de classements (*Indices et Classements*), deux bases de textes, l'une sur la *Concurrence*, l'autre sur des thèmes plus généraux (*Intradoc*), et enfin un dispositif spécialisé en information de nature juridique (portail *Actualités juridiques*)[8].

Dans l'espace informationnel ainsi structuré, l'interrogation en langage naturel qui sera étudiée dans la suite de cet article porte uniquement sur les bases textuelles *Intradoc* et *Concurrence*.

Des modalités d'interrogation diversifiées

L'une des conséquences les plus importantes, nous semble-t-il, des diverses enquêtes menées auprès des usagers a été d'abandonner l'idée de trouver le « meilleur » (sous-entendu : « et unique ») mode d'accès à l'information. Des modes d'accès par plan de classement, par mots clés, en langage naturel, par champs structurés – chacun a son utilité en fonction du profil de l'usager, de son besoin, du contexte de sa recherche, des types d'information et de documents proposés.

Aussi, en complément de la recherche « libre », c'est-à-dire sans contrainte de désignation d'un champ ni utilisation d'une syntaxe particulière, un effort particulier a été fait concernant l'interrogation par thèmes et par champs structurés (titre du document, titre de la revue, auteur, type de documents, nom de société ou espace géographique). Ces différentes modalités d'interrogation sont pleinement utilisées, comme le montre le tableau de la page 322.

2 Analyse des questions posées et de l'utilisation du SID

Depuis un an, le responsable du Centre documentaire des AGF reçoit et traite mensuellement le fichier contenant la liste des questions posées (voir pages 318-319). L'étude de deux mois (août et octobre 2000, soit environ 4.500 questions posées par près de 1.100 personnes différentes) nous a permis de mettre en évidence un certain nombre d'aspects positifs du site *Ressources documentaires* ainsi que les difficultés qui se posent toujours aux usagers. Certains des points évoqués sont spécifiques aux dispositifs en langage naturel, d'autres relèvent de problématiques plus générales de recherche documentaire. Il nous a paru intéressant de les traiter ensemble.

Les différents résultats devront bien sûr être confirmés, affinés et/ou étendus par l'étude ultérieure des questions posées sur d'autres périodes.

Le nombre d'usagers est en forte augmentation

Plus que l'augmentation du nombre de recherches documentaires, c'est d'une part le nombre et la diversité des usagers différents (460 en août, 638 en octobre), et d'autre part l'utilisation régulière du système qui renforcent le sentiment que le Centre documentaire a fait, en 1998, un choix pertinent. En effet, depuis l'ouverture de l'intranet documentaire, avec des modalités de recherche simplifiées et un accès direct à l'information, le nombre de consultations progresse régulièrement. Le tableau de la page 322 donne

8 Le site Ressources documentaires *propose également en ligne d'autres informations comme les archives (un lien avec le site des Archives historiques), les sites web sélectionnés et commentés, les publications (des notes internes sur le marché des assurances, des circulaires d'organismes officiels, etc.).*

des indications complémentaires sur le nombre de questions posées par usager[9].

Bien sûr les données chiffrées, même si elles sont importantes, ne suffisent pas pour évaluer la pertinence et l'efficacité du dispositif. En plus d'autres enquêtes qui seront menées ultérieurement[10], l'analyse plus précise des questions posées et des modalités de formulation des requêtes doit nous permettre, comme nous allons le voir à présent, de mieux comprendre les usagers lors de leur activité de recherche documentaire, et leurs difficultés.

L'exploitation d'une base peut s'avérer inappropriée à certaines questions

Certaines questions posées montrent que les usagers éprouvent des difficultés à se faire une image de l'espace informationnel, à se représenter le contenu de ce qui leur est proposé (voir page 315, étape 1) – contenu bien connu, par contre, des documentalistes puisque construit par eux[11] ! Cette difficulté de compréhension du périmètre du dispositif et de sa composition peut être résolue, comme nous venons de le voir, d'une part grâce à un découpage de l'espace informationnel intelligible par les usagers, parce que conçu à partir de leurs pratiques, et d'autre part grâce au choix d'une terminologie adaptée.

Il restera toujours des questions qui relèveront à la fois de l'une et de l'autre des différentes parties de l'espace informationnel, rendant plus délicate l'orientation sur le site (« le marché de tel produit »), voire des questions qui resteront sans réponse faute... de données[12]. Mais l'objectif est de donner à l'usager une vision plus pertinente du site, lui permettant de mieux appréhender le dispositif qu'il va utiliser. Le travail fait sur l'organisation du site et sur sa lisibilité (terminologie employée, organisation spatiale sur les écrans, etc.) constituent une partie de la réponse à ces difficultés.

Dans un système en langage naturel, il faut s'exprimer en langage naturel

Pour fonctionner efficacement, un système en langage naturel doit pouvoir s'appuyer sur une

9 Malgré l'impossibilité d'identifier nommément les usagers, le système distingue les utilisateurs entre eux grâce à l'identifiant de l'ordinateur connecté.

10 Un premier sondage a permis d'identifier certains critères d'utilisation du site (ou du dispositif) par les usagers : accessible à n'importe quelle heure, dans son environnement de travail ; en toute liberté ; pour « voir ce qu'il y a »...

11 Aux AGF, les documentalistes sélectionnent, numérisent et traitent les documents.

12 Par exemple, les documents internes sont absents du fonds documentaire, en particulier ceux produits par la direction des ressources humaines ou de la direction juridique.

Éléments bibliographiques

• Dimension « usager »

[1] Le besoin d'information : formulation, négociation, diagnostic / Yves LE COADIC. – Paris : ADBS Éditions, 1998 [Introduction générale à l'étude du besoin d'information exprimé par une question]

[2] Comment les logiciels de bases de données bibliographiques et textuelles peuvent-ils répondre aux différents besoins de leur utilisateurs ? [en ligne] / Suzanne BERTRAND-GASTALDY. – [Consulté le 28 août 2000]. – <http://www.ling.uqam.ca/sato/publications/bibliographie/Ind_lang.htm>

[3] Vers une ergonomie linguistique [dossier] / Yolla POLITY. – *Documentaliste - Sciences de l'information*, 1994, vol. 31, n° 3, p. 135-158

• Dimension « Indexation, texte intégral, traitements linguistiques et recherche en langage naturel »

[4] Quelques liens sur le TALN et la linguistique [en ligne] / TALANA. – [Consulté le 2 janvier 2001]. – <http://talana.linguist.jussieu.fr/talwww.html>

[5] À la découverte de l'ingénierie linguistique en France [en ligne] / Délégation générale à la langue française. ([Consulté le 2 janvier 2001]. – <http://www.culture.fr/culture/dglf/riofil/garde.htm>

[6] Linguistique : domestiquer l'ordinateur [dossier]. - *Archimag*, novembre 2000, n° 139

[7] Indexation automatique, recherche d'information et évaluation / Pierre LE LOARER. - *In* : Le traitement électronique du document, cours INRIA, 2-7 octobre 1994, Aix-en-Provence, coord. par Jean-Claude Le Moal et Bernard Hidoine. - Paris : ADBS Éditions, 1994. – P. 149-201

[8] 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, cours INRIA, 30 septembre-4 octobre 1996, Trégastel, coord. par Jean-Claude Le Moal et Bernard Hidoine. - Paris : ADBS Éditions, 1996. – P. 127-164

[9] Moteurs d'indexation et de recherche / Catherine LELOUP. – Paris : Eyrolles, 1997

[10] Recherche documentaire : du thésaurus au texte intégral / Philippe Lefèvre. – Paris : Hermès, 2000

[11] Documentation commerciale de logiciels en langage naturel (Spirit de T.GID, Lexiquest, Intuition, Micro-Mind de Sagitex, etc.)

• Méthodes d'évaluation

[12] TREC (Text Retrieval Conference) : une conférence pour l'évaluation des systèmes d'information / Karine LESPINASSE. – *Documentaliste - Sciences de l'information,* 1997, vol. 34, n° 2, p. 74-81

[13] Amaryllis [en ligne]. – [Consulté le 2 janvier 2001]. <http://www.inist.fr/accueil/profran.htm>

• Conception de mémoires documentaires

[14] La recherche d'information dans les mémoires électroniques : l'enjeu documentaire / Hubert FONDIN. – *Documentaliste - Sciences de l'information,* 1999, vol. 36, n° 4-5, p. 242-248 [Le façonnage préalable des « objets informationnels dans les mémoires électroniques » constitue l'enjeu documentaire aujourd'hui. Comparaison des problèmes de recherche documentaire entre les périodes 1957-70 et 1994-98]

phrase ou au moins une expression. Les pratiques d'interrogation qui consistent à « taper quelques mots » pour exprimer un besoin documentaire, sans substantif ou sans verbe (certains usagers sont même atteints du syndrome du booléen anglais, comme dans les questions « *agf and délais de publication des comptes* » ou « *gestion de crise and perte d'image* », par exemple !), réduisent considérablement l'efficacité des systèmes linguistiques, parce que les traitements linguistiques ne peuvent pas être mis en œuvre. On se retrouve alors dans le cas d'un traitement en « texte intégral ».

Il semble que cette pratique ait un effet négatif, bien sûr sur le nombre et la pertinence des réponses (elle génère à la fois du silence et du bruit), mais surtout sur la présentation des résultats en « classes » (leur contenu, leur ordre). En effet, les logiciels en langage naturel recourent à des techniques statistiques[13] pour pondérer les termes d'indexation à partir des résultats des traitements linguistiques ; lorsque ceux-ci sont peu, mal ou pas du tout mis en œuvre, les résultats des traitements statistiques sont en conséquence dégradés.

13 Mises en œuvre dans les logiciels en « texte intégral » tels Search 97 ou Fulcrum, par exemple.

Sans aller jusqu'à utiliser la forme du dialogue (voir ci-contre la liste de questions), il est important de s'exprimer librement et de la manière la plus proche possible du langage naturel.

Les traitements sémantiques restent insuffisants

Certaines questions posées nécessiteraient des traitements sémantiques de plus haut niveau que ceux proposés aujourd'hui par la plupart des logiciels en langage naturel. Par exemple, les questions qui appellent des réponses de type oui/non ; ou bien encore lorsque l'usager ne souhaite pas de documents en réponse à une question, mais directement une réponse à un problème.

Dans ce cas de figure, l'ensemble des termes utilisés dans l'expression de la question et la syntaxe employée ne se trouvent pas nécessairement dans le contenu des textes interrogés ; la reformulation proposée par le logiciel à l'aide de traitements morphosyntaxiques ne suffit plus. Exemples : « *définition de la subrogation légale* », « *définition du conjoint* », « *les démarches à réaliser pour le mariage* », « *qu'est-ce qu'une stratégie ?* », « *qu'est-ce qu'un agent de maîtrise ?* », « *comment assurer une tondeuse, un scooter ?* », ou encore « *peut-on adresser au client la copie du rapport d'expertise médicale ?* », « *quelle est la position de la jurisprudence en assurance de personne sur le suicide de l'assuré ?* »,

Exemples tirés du fichier des questions posées sur l'intranet documentaire en octobre 2000

N°	N° des postes utilisateurs	Base interrogée	Termes des questions posées[1]		
12	128.193.18.68	Intradoc	ASSURANCES DOMMAGES[2]	[`01/10/1999`, `31/12/2000`][3]	
128	128.193.225.189	Intradoc	Ouvrage	réassurance	ramel
223	128.193.244.88	Intradoc	indemnités journalières pour commerçants		
316	128.65.224.217	Intradoc	assurance persistance de frais généraux		
346	126.65.225.137	Intradoc	une tempête peut-elle être exonératoire de toute responsabilité ?		
419	128.65.225.97	Intradoc	relations sociales	Article	ENTREPRISE
606	129.129.224.166	Concurrence	AXA Courtage	Coralis	
745	129.65.225.67	Intradoc	quelle est la position de la jurisprudence en assurance de personne sur le suicide de l'assuré ?		
1691	133.194.225.83	Intradoc	les plates-formes téléphoniques pour l'indemnisation des sinistres		
1698	133.66.116.105	Intradoc	plateforme agf		
1725	133.66.224.145	Concurrence	PRODUITS	WINTERTHUR	
1726	133.66.224.145	Concurrence	RESEAU DE DISTRIBUTION	WINTERTHUR	
1766	133.66.224.145	Intradoc	expatriés	Article	Agefi

1 Termes des questions posées : soit en langage naturel sur la zone « libre », soit dans un ou plusieurs champs structurés (exemples : Auteur, Domaine, Société, Type de document).

2 En majuscules (ASSURANCE DOMMAGES, ENTREPRISE, etc.) : sélection d'un terme dans une liste thématique (10 domaines principaux, suivis de 22 sous-domaines), ou nom de société (WINTERTHUR).

3 [`01/10/1999`, `31/12/2000`] : code généré automatiquement par le système lors de recherches thématiques pour permettre la visualisation des documents les plus récents.

Quelques exemples de questions posées

Voici quelques formulations de questions posées dans la zone de recherche en langage naturel (zone libre) :

« offre d'assurance en IART pour la boulangerie »

« quelles mesures prendre en matière de santé pour continuer à être couvert en cas de maladie lors d'un voyage à l'étranger ? »

« menace grave imminente d'effondrement »

« faute inexcusable en Badinter »

« les liens capitalistiques entre banques et assurances »

« la convention internationale de partage RC/tierce »

« taux de conversion de rente loi Madelin »

« Merci de me communiquer toute information concernant GAIPARE SELECTION »

« le mur de soutènement est-il garanti par un contrat d'habitation »

« définition de la subrogation légale »

« Définition du conjoint »

« Les démarches à réaliser pour le mariage »

« qu'est-ce qu'une stratégie ? »

« qu'est-ce qu'un agent de maîtrise ? »

« comment assurer une tondeuse, un scooter ? »

« peut-on adresser au client la copie du rapport d'expertise médicale ? »

« quelle est la position de la jurisprudence en assurance de personne sur le suicide de l'assuré ? »

« naissance, évolution et mort d'une start-up »

Réponses à des questions posées sous Spirit dans l'intranet documentaire des AGF

Questions / réponses sur Intradoc (14.724 documents sur tous sujets)
Présentation par « classes » des réponses

Quelle est la définition de l'enfant à charge ?

CLASSE 1	enfant-charge, définition dont un chapitre d'un ouvrage de droit et protection sociale	23 réponses
CLASSE 2	enfant-charge	42 réponses
CLASSE 3	enfant, charge, définition	250 réponses

Ici, les termes *enfant-charge* sont reliés syntaxiquement dans les documents comme demandé dans la question, le terme « définition » étant par contre plus éloigné dans le texte des documents. La deuxième classe de documents ne comporte pas le terme *définition* ; la troisième regroupe des textes pour lesquels les trois termes (*enfant, charge* et *définition*) ne sont pas reliés syntaxiquement, donc *a priori* d'un intérêt bien moindre, voire sans intérêt.

Difficultés juridiques relatives à la vente en ligne de contrats

CLASSE 1	difficultés-juridiques-relatives-vente-ligne-contrat	1 réponse
CLASSE 2	juridiques-relatives, vente-ligne, difficultés, contrat	1 réponse
CLASSE 3	vente-ligne-contrat, difficultés, relative	2 réponses
CLASSE 4	vente-ligne-contrat, juridiques, relative etc.	1 réponse

Le document unique de la première classe contient l'ensemble des termes reliés syntaxiquement comme dans la question.

Comment s'assurer contre un accident de tondeuse à gazon ?

CLASSE 1	accident-tondeuse	1 réponse
CLASSE 2	assurer, accident, tondeuse, gazon	5 réponses
CLASSE 3	accident, tondeuse, gazon	1 réponse
CLASSE 4	assurer, tondeuse, gazon	1 réponse
CLASSE 5	assurer, accident, tondeuse	1 réponse

La première réponse propose un article de *Que choisir ?* d'octobre 1999 sur les accidents de tondeuse. La deuxième classe, la plus intéressante, propose cinq articles dont un, de juin 2000, sur les nouvelles garanties pour les accidents de la vie (« une personne se blesse en tondant sa pelouse »). On notera la reformulation (tondeuse à gazon => tondant sa pelouse).

Le mur de soutènement est-il garanti par un contrat d'habitation ?

CLASSE 1	mur-soutènement, garanti, contrat, habitation	2 réponses
CLASSE 2	contrat-habitation, mur, garanti	2 réponses
CLASSE 3	mur-soutènement, contrat, habitation	1 réponse
CLASSE 4	mur-soutènement, habitation	1 réponse

Taux de conversion de rente loi Madelin

CLASSE 1	taux-conversion-rente-loi-madelin	1 réponse
CLASSE 2	conversion-rente, loi-madelin, taux	2 réponses
CLASSE 3	taux-conversion, loi-madelin	1 réponse
CLASSE 4	loi, rente	111 réponses
CLASSE 5	taux, rente	93 réponses
CLASSE 6	madelin	14 réponses
CLASSE 7	conversion	150 réponses

« *naissance, évolution et mort d'une start-up* ».

Pour comprendre les problèmes qui se posent ici, il faut reprendre la démarche méthodologique de recherche documentaire. Celle-ci propose classiquement une procédure en plusieurs grandes étapes :

- l'analyse du problème ;
- le repérage et la sélection d'un (ou de plusieurs) type(s) de ressources en fonction de la question : un dictionnaire ou une encyclopédie spécialisée pour une recherche de définition, un texte de loi ou de la jurisprudence, des rapports ou études pour un état de l'art sur la vente en ligne, un compte rendu d'activité d'une société concurrente, ou encore un contrat-type pour traiter un genre de contrat non encore pris en compte dans une agence (accidents liés aux tondeuses ou aux quads[14], par exemple) ;
- la localisation sur l'intranet de la ressource qui contiendrait la réponse : les indicateurs (*Indices* et *Classements*) pour des données chiffrées sur l'indice FFB ou sur le positionnement de tel concurrent, la base *Intradoc* pour une étude sur le marché des assurances, la base *Concurrence* pour des articles sur les produits d'assurance d'un concurrent ;

Profils d'usagers et types d'usages

L'étude des questions posées au Centre documentaire des AGF sur l'intranet documentaire permet de repérer des familles différentes d'usagers (selon les types de documents recherchés, les domaines de recherche) et d'usages

Familles d'usagers	Besoins liés directement aux activités professionnelles
Juriste et « paralegals »	Au moment de l'élaboration d'un produit (en amont), ou de son usage (en aval, vis-à-vis du problème d'un client)
Marketing	Produits, marché, techniques de marketing, promotion, communication
Stratégie et planification	Pour réaliser des notes, études de marché, économiques et macro-économiques
Concepteur de formation et formateurs	Sur les métiers, les techniques de formation, l'animation de stages, voire la création de modules
Gestionnaire de dossiers d'assurance	Sur des situations d'assurance, pour des particuliers ou des entreprises (litiges, contentieux, etc.)
Stagiaire	Pour des informations générales sur l'assurance, le secteur, les modes de distribution, les acteurs du marché
Tout collaborateur	Sur le « salarié dans son entreprise » (rémunération, RTT, entretien d'appréciation, retraite, etc.)

Types d'usages

Retrouver une information : « listes des agences en gestion plates-formes téléphoniques », « taux de conversion de la rente loi madelin »
Retrouver un document : « Ouvrage / CAPA / seniors », « Article / assureurs retrouvent le sourire / Argus »
Formuler, pour un problème clair, explicité et bien identifié pour l'utilisateur, une question appropriée, soit sur un thème, soit sur une caractéristique de l'information ou du document (tels un type de document, une période, etc.) : « Produits / Carrefour / opportunités d'épargne »
Liberté de fouiller, de butiner : recherche sur un thème parmi ceux proposés, recherche d'une société dans la base Concurrence, etc.
Faire des tests sur le système pour le comprendre

Lors d'une classification des métiers exercés aux AGF en 1996, la DRH avait défini plus d'une centaine de familles que l'on retrouve dans ces familles d'usagers, sans qu'elles se recouvrent complètement. Certains « métiers » sont plus demandeurs d'information et de documents.

- enfin une (ou plusieurs) recherche(s) d'information et/ou de documents, à l'intérieur de la ressource retenue, en fonction de modalités propres à la ressource utilisée (langage naturel, menu hiérarchique, etc.).

Le cas d'une recherche de « définition ». La réponse attendue par l'usager pour la définition d'un concept (« qu'est-ce que la stratégie ? » ou « qu'est-ce qu'un agent de maîtrise ? ») peut se trouver dans un article qui traiterait de méthodes en stratégie ou du développement de l'emploi chez les agents de maîtrise. Un auteur peut, en introduction, présenter des définitions des concepts traités dans le texte. La valeur attribuée à ce type de ressource (un article plutôt qu'un dictionnaire, même spécialisé) ne peut être établie que par l'usager lui-même. C'est pourquoi la recherche (la fouille, devrait-on dire) de l'information à partir de la lecture de quelques documents proposés par le système est apprécié par les lecteurs, et peut se révéler très riche pour eux. Le problème réside alors d'une part dans la fourniture de « quelques documents » en nombre raisonnable et d'autre part dans des possibilités de consultation et d'exploitation aisées de ces documents (voir ci-après).

Le cas des questions multiples. L'analyse de certaines demandes (« *naissance, évolution et mort d'une start-up* », par exemple) renvoie à plusieurs questions qui concernent de multiples aspects documentaires. Dans l'exemple cité, plusieurs documents pourront répondre à la question, les uns sur le développement, d'autres sur la disparition des *start-up*.

Dans les interfaces de recherche booléenne, il est possible d'élaborer des requêtes complexes en utilisant des parenthèses et des opérateurs divers. Dans le cas d'une recherche en langage naturel, les réponses seront réparties dans plusieurs classes, certaines renvoyant à des documents ne traitant qu'une partie de la question posée. Mais c'est à l'usager d'aller « piocher » dans les différentes classes pour y trouver son bonheur à partir de documents prenant plus ou moins en compte la question.

Des traitements sémantiques de haut niveau seraient nécessaires ici, ce type de question générant simultanément beaucoup de bruit et de silence. Actuellement, le niveau des traitements proposés ne permet pas d'envisager des solutions opérationnelles totalement automatiques pour ces questions multiples.

Les résultats sont organisés par lots et pondérés

Les documents ne sont pas égaux devant la question posée : l'ensemble de ceux que l'on pourra obtenir en réponse à une question ne seront pas tous intéressants, voire pertinents. Dans les systèmes classiques, aucune pondération du lot résultat n'est effectuée ; chaque document est estimé pertinent. C'est la loi du tout ou rien. Dans un dispositif en langage naturel sous Spirit, l'organisation des résultats par « classes » (voir page 315, étape 4), avec la pondération et le tri automatiques que cela suppose, permet à l'utilisateur de s'orienter : on lui propose d'abord un choix parmi plusieurs groupes de documents, rassemblés en raison de leur degré d'appariement plus ou moins fort à la question, puis, dans un deuxième temps, le choix d'un document au sein de l'un de ces groupes.

Le système assiste l'utilisateur dans la fouille au sein d'une classe et dans le butinage au sein des documents en lui offrant des possibilités de circulation à l'intérieur d'un document ou entre plusieurs d'entre eux, en passant « d'information en information », c'est-à-dire de l'un à l'autre des groupes de termes utilisés à la recherche.

La langue est vivante

Les outils standards de traitement en langage naturel ne résolvent pas, dans la traduction d'un concept, tous les problèmes de terminologie et d'expressions multiples. L'ambiguïté de certains noms propres (des noms de société comme « W Finances »), en particulier, ou les différentes formes que peut prendre un même terme (« A.G.I.R.A. » et « AGIRA », ou encore « EAA » pour « entretien annuel d'appréciation », par exemple) peuvent générer du silence. La reformulation reposant sur le dictionnaire de la langue, de même que les règles syntaxiques et grammaticales s'avèrent parfois insuffisantes, comme dans le cas des termes construits à partir de « télé » ou « cyber ». Spirit trouvera bien « cyberconsommateur », mais pas son équivalent « internaute consommateur », car celui-ci est actuellement absent du dictionnaire privé. De même, les mots commençant par « télé » seront repérés s'ils sont écrits sans séparateur (« télé-acteur » n'est pas reconnu).

Concernant les différentes formes d'un même terme, en particulier pour les sigles et leur développement, il est important de prendre la mesure de la logique de recherche sur le contenu. Celle-ci repose sur le constat de la présence, dans un même document, de plusieurs formulations différentes d'un même concept. Par exemple, l'un et l'autre des termes suivants : « EAA » et « entretien annuel d'appréciation », ou encore « entretien d'appréciation annuel » et « entretien d'évaluation annuel », peuvent coexister dans un même article de presse.

Aussi d'autres solutions doivent-elles être mises en œuvre. L'une, employée essentiellement pour les noms propres, consiste à définir des champs factuels (société, produits) que les documentalistes

14 Quad : engin motorisé.

LES PRATIQUES DES USAGERS ET LEUR ÉVOLUTION

Les questions posées au cours des mois d'août et octobre 2000 ont été triées et étudiées. Un premier niveau d'analyse, présenté ici, permet de proposer des critères pour l'établissement d'un tableau de bord ; celui-ci sera consolidé lors de l'étude des questions posées par les usagers dans les mois à venir.

Rappel : en août 2000, 460 usagers ont posé 2.242 questions, dont 407 sur la base *Concurrence* et 1.835 sur la base *Intradoc* ; en octobre 2000, 638 usagers ont posé 2.735 questions, dont 360 sur la base *Concurrence* et 2.375 sur la base *Intradoc*.

En effectuant une comparaison entre les résultats des mois d'août et octobre (colonne de droite), on fait apparaître des évolutions particulièrement significatives.[1]

Fréquence d'utilisation : nombre de questions posées par usager

	AOUT 2000		OCTOBRE 2000		ÉVOLUTION
Nombre de questions posées par usager	Nombre d'usagers	Pourcentage d'usagers	Nombre d'usagers	Pourcentage d'usagers	ENTRE AOÛT ET OCTOBRE
1	170	36,96 %	217	34,01 %	+ 27,65 %
2	87	18,91 %	133	20,85 %	+ 52,87 %
3	44	9,57 %	73	11,44 %	+ 65,91 %
4	44	9,57 %	49	7,68 %	+ 11,36 %
5	18	3,91 %	33	5,17 %	+ 83,33 %
6 à 10	47	10,22 %	84	13,17 %	+78,72 %
11 à 50	46	10 %	47	7,21 %	–
> 50	4	0,87 %	3	0,47 %	– 25,00 %
Total	460	100 %	638	100 %	+ 38,70 %

Nombre de critères de recherche utilisés

	AOUT 2000		OCTOBRE 2000		ÉVOLUTION
Nombre de critères de recherche utilisés par question posée	Nombre de questions posées	Pourcentage de questions posées	Nombre de questions posées	Pourcentage de questions posées	ENTRE AOÛT ET OCTOBRE
1	1.329	59,28 %	1.603	58,61 %	+ 20,62 %
2	687	30,64 %	892	32,61 %	+ 29,84 %
3	215	9,59 %	178	6,51 %	– 17,21 %
4	11	0,49 %	62	2,27 %	+ 463,64 %**
Total	2.242	100 %	2.735	100 %	+ 21,99 %

Types de critères de recherche utilisés

	AOUT 2000		OCTOBRE 2000		ÉVOLUTION
Types de critères de recherche utilisés	Nombre de questions posées	Pourcentage de questions posées	Nombre de questions posées	Pourcentage de questions posées	ENTRE AOÛT ET OCTOBRE
Types de documents (ouvrages, articles, etc.)	302	10,23 %	411	8,77 %	+ 36,09 %
Titres de revues	74	2,51 %	74	1,58 %	–
Un thème	1.714	58,08 %	2.583	55,10 %	+ 50,70 %
- dont thème hors liste proposée	1.294	43,85 %	1.373	29,29 %	+ 6,11 %
- dont un thème choisi dans la liste proposée	420	14,23 %	1.210	25,81 %	+ 188,10 %*
Noms de sociétés ou de produits	441	14,94 %	410	8,75 %	– 7,03 %
Total	2.951	100 %	4.688	100 %	+ 58,86 %

[1] *L'analyse de ces chiffres doit être approfonde. Par exemple, l'augmentation de la recherche sur un thème choisi dans la liste proposée (*) doit être mise en relation avec celle de la recherche multicritères (**). Tableau élaboré avec la participation de Joëlle Cohen.*

alimentent (très peu il est vrai, dans le cas des AGF) avec une sélection de noms de sociétés jugées les plus importantes, par exemple. La recherche est ainsi optimisée (en particulier en évitant les risques d'erreur orthographique), sans exclure toutefois la recherche sur les parties textuelles dans le cas où il s'agirait d'une nouvelle société ou d'une société d'importance faible. Cette technique n'est évidemment pas spécifique aux outils en langage naturel.

L'autre solution proposée par Spirit, principalement pour les noms communs, repose sur le développement d'un dictionnaire privé, en complément du dictionnaire général de la langue. Ce dictionnaire privé sera utilisé soit pour les termes et expressions relevant du métier de l'entreprise (« réassurance », « assurance de dommages », « assurance IART » ou « coassurance »), soit en raison de pratiques linguistiques particulières (anglicisme comme « *embedded value* »). Il est utilisé automatiquement avant le dictionnaire général et il est mis à jour par le Centre documentaire des AGF. Une centaine de termes y ont été à ce jour intégrés, avec leurs différentes formes comme « bonus malus », « boni mali », « bonus-malus », ou encore « AXA » pour éviter la confusion avec le verbe « axer ».

Les erreurs de frappe (ou fautes d'orthographe)

Des erreurs de saisie ou d'orthographe sont une autre cause de dysfonctionnements : sur 2.044 questions posées en août 2000 par environ 500 utilisateurs, 25 comportaient des erreurs de frappe ou d'orthographe, comme par exemple « *acident* » ou encore « *indemnites joournalieres pour commerçants* ».

Les dispositifs fondés sur le texte intégral traitent la question en prenant la chaîne de caractères telle qu'elle est écrite par l'usager. Seuls les systèmes ayant une « intelligence » linguistique permettent de corriger ces erreurs, soit en indiquant à l'usager le mot inconnu, à charge pour lui d'effectuer la correction, soit en le traitant automatiquement.

Dans sa version générale, Spirit inscrit dans un écran particulier les termes qui lui sont inconnus. Sur l'intranet documentaire des AGF, il a semblé préférable de simplifier l'interface utilisateur. Aussi un message[15] apparaît seulement lorsqu'il n'y a aucune réponse à la question posée, ce qui peut être dû, bien sûr, à une toute autre raison qu'une erreur de frappe. Ce point pourrait être étudié ultérieurement.

15 « Aucune réponse. Merci de bien vouloir contacter le Centre documentaire des AGF. ».
16 Depuis plusieurs mois, les statistiques montrent que le nombre des questions posées se situe entre 2.200 et 2.500 pour environ 800 à 1.000 intranautes réguliers.

3 Tous les problèmes documentaires ne sont pas résolus par le langage naturel

Si le développement de l'utilisation du fonds documentaire (de deux cents usagers en 1995 sous Basis à environ un millier en 2000[16]) montre un succès certain de cette nouvelle formule, l'option « langage naturel » ne résout pas tous les problèmes posés à l'ensemble des usagers (finals et documentalistes) par la recherche documentaire. Des efforts restent à faire : de nature technologique (amélioration des traitements statistiques et/ou linguistiques) ou terminologique (gestion facilitée du dictionnaire privé), qui sont du ressort de l'éditeur du logiciel ; ou encore de nature culturelle, par le développement au sein des AGF d'une meilleure approche de l'information et de la recherche d'information.

L'amélioration de l'exploitation des systèmes documentaires par l'usager final n'est pas exclusivement due, dans le cas des AGF, au choix du langage naturel : le passage à un nouveau logiciel a été l'occasion de repenser globalement le dispositif, l'organisation de l'« espace informationnel », les modalités et l'ergonomie de l'accès à l'information, en centrant réflexion et choix sur les besoins et surtout les pratiques des usagers.

Au terme de cette étude, non systématique et portant sur une durée courte, des questions posées sur les bases documentaires accessibles en langage naturel, plusieurs remarques générales peuvent être formulées.

Nous nous éloignons d'une logique du « tout ou rien »

Mots clés contre langage naturel, automatique contre manuel, base de références contre base de textes, etc. La conception d'un dispositif documentaire doit être perçue comme un travail créatif, orientée « exploitation directe par des usagers ». Un tel dispositif s'appuie d'une part sur la construction d'un espace où doivent se rencontrer des usagers d'information et des informations, le plus efficacement et le plus simplement possible ; d'autre part sur la coordination et la mise en cohérence de nombreux outils documentaires et informatiques (index, dictionnaire, indexation manuelle et automatique, champs structurés, zone de saisie, ergonomie, terminologie et fonctions proposées, etc.).

Nos métiers évoluent

Les activités des documentalistes évoluent d'une activité de construction d'une base de don-

nées à des activités d'ingénierie documentaire, elles se transforment, s'enrichissent, se complexifient :

- d'un travail régulier, voire routinier (indexer des documents et alimenter un fichier de références), le traitement des documents diminue globalement au profit d'un travail, effectué en amont, de sélection[17] et d'analyse (extraction de données et enrichissement de bases d'informations spécialisées, classification, plus forte structuration de l'analyse, etc.) ;

- les tâches d'administration du vocabulaire se réduisent en charge et se complexifient, le travail sur un dictionnaire et sur les traitements linguistiques n'étant pas de même nature ni de même niveau que celui lié à la maintenance d'un thésaurus ;

- en choisissant de placer l'usager au centre du dispositif, la conception, l'organisation et l'ergonomie de ce dernier deviennent des activités fondamentales, renouvelées pour prendre en compte les technologies de l'information les plus récentes.

La question de la double compétence des documentalistes

Un autre intérêt que présente pour les documentalistes un dispositif en langage naturel vient de ce qu'il est possible de s'affranchir, dans une certaine mesure et dans le cas d'un centre de documentation généraliste, de la spécialisation des fonds documentaires.

La documentation a été initiée au début du siècle dans le domaine de la recherche (fondamentale ou industrielle). Beaucoup de documentalistes étaient porteurs d'une double compétence, documentaire et scientifique ; bien souvent seules les compétences scientifiques étaient demandées, les autres étant acquises « sur le tas ». Ce profil professionnel était nécessaire et il s'est déployé dans d'autres secteurs que celui des sciences « dures » : en économie, dans les secteurs de la finance et du droit... Ce fonctionnement est très exigeant, puisqu'il suppose une formation permanente des documentalistes dans ces domaines (ce qui n'est pas toujours le cas, il est vrai). Si cela se justifie dans certains environnements où le documentaliste a une mission réelle d'appui et d'assistance de proximité auprès des usagers, cela se justifie moins dans le cas du centre documentaire des AGF, dont la mission est de mettre en relation des utilisateurs et les ressources informationnelles externes, et pour lequel la gestion des flux d'information est jugée prioritaire.

En effet, les documentalistes de ce centre, en dehors des documentalistes-juristes, ne sont pas nécessairement spécialisés, ni en ressources humaines, ni en finances, domaines mal maîtrisés où se posaient des problèmes d'indexation. La compréhension du domaine traité s'opère à un niveau général (suivre les grands acteurs, les grands moments ou étapes du secteur, ponctuellement se pencher sur une thématique importante, nouvelle, etc.), ce qui rend le travail documentaire plus léger et moins contraignant par rapport aux savoirs et savoir-faire du domaine.

4 Une première étude à renouveler

Au terme de cette première étude de l'utilisation de l'intranet documentaire, le Centre documentaire des AGF va poursuivre sa surveillance des besoins des usagers et de l'usage du dispositif. L'attention de l'équipe portera bien sûr sur la recherche en langage naturel. [...]

Ces premiers résultats positifs ne doivent pas nous faire oublier que tous les utilisateurs potentiels du centre ne sont pas encore concernés par l'intranet documentaire, que tous les besoins des usagers actuels (voir page 320) ne sont pas forcément couverts, et que d'autres ressources spécialisées au sein des AGF pourraient être mutualisées avec profit pour la collectivité.

Ainsi, un projet nécessitant la reconception visuelle de la page d'accueil est en cours de réflexion ; ce projet sera encore plus nécessaire si d'autres centres documentaires rejoignent l'intranet.

D'autres projets encore sont prévus pour 2001, comme le développement d'une base *Bilans de sociétés*, de la base FAQ, d'une meilleure signalisation d'études acquises et d'un dossier documentaire électronique. Ce dossier, portant sur un thème précis et conservé en ligne pendant un mois, serait constitué d'extraits du site complétés par des liens vers d'autres sites. Enfin, des accès au site *Ressources documentaires* par l'intermédiaire d'extranets sont en cours de réalisation pour les courtiers, agents et inspecteurs d'assurance.

DÉCEMBRE 2000

17 La politique de sélection du Centre documentaire des AGF, qui n'a pas une vocation patrimoniale sur les documents externes, est devenue plus sévère et plus rigoureuse en raison d'une part des coûts de l'information électronique et d'autre part du peu d'intérêt de traiter à long terme certains documents non pérennes.