TECHNO 18

Publication technique de la SmalS-MvM / Technische uitgave van de SmalS-MvM

# Les logiciels de recherche documentaire

*Françoise Vanden Bossche est chef de projets GED à la SmalS-MvM depuis 1997. Depuis 1984, elle a travaillé pour diverses institutions publiques (parlements belge, wallon, germanophone, luxembourgeois) et pour des entreprises privées à l'élaboration de bases juridiques, de revues de presse ou de bases d'information.*

*Contact : 02/509.59.78*

## 1. Introduction

logiciels de recherche documentaire servent à retrouver une information dans une masse de textes. Bon nombre de sociétés utilisent ces logiciels pour gérer leur documentation : textes juridiques, réglementaires ou expliquant les procédures de travail, revues de presse, photothèques, documents concernant un sujet particulier, ...

Ces logiciels permettent une gestion évoluée des centres de documentation et, de façon plus générale, effectuent des recherches dans un ensemble de documents.

La SmalS-MvM a acquis le logiciel Basisplus pour gérer ses applications documentaires. Deux applications sont déjà en production à l'ONSS et à l'ONSS-APL. Elles permettent d'effectuer des recherches dans la documentation juridique et administrative des instances directoriales.

Les logiciels documentaires sont apparus sur le marché à la fin des années 50. Leur vocation était de permettre d'effectuer des recherches non définies au préalable dans des masses d'informations, constituées de collections de textes libres – du style de document en traitement de texte.

Jusqu'à la deuxième moitié des années 80, ces logiciels permettaient uniquement d'effectuer des recherches: toute autre fonctionnalité requise demandait un interfaçage avec d'autres logiciels, entre autres pour introduire les données, pour les imprimer, pour les charger dans la base, ...

Nous verrons l'évolution de ces logiciels quant à **l'interfaçage avec les logiciels bureautiques**.

En outre, il n'était pas possible d'établir des **liens entre documents** – par exemple entre différentes versions d'un document. Nous verrons que les documents peuvent également être retrouvés, sur base de leur lien avec d'autres documents.

Depuis le boom Internet, les logiciels documentaires connaissent un nouvel essor. Ils guident l'internaute à libeller ses recherches dans une base dont il ne connaît ni la structure, ni le contenu.

Comme les bases consultables sur Internet, les bases documentaires doivent pouvoir être utilisées par des **utilisateurs non avertis** et dont les besoins en consultation peuvent être des plus divers.

Cette première constatation impose aux concepteurs d'applications documentaires un principe de base essentiel, à savoir la **convivialité des interfaces** et la recherche d'une utilisation intuitive.

Dans ce document, nous présenterons d'abord quelques particularités des logiciels documentaires et nous développerons ensuite les possibilités fonctionnelles offertes par ces logiciels pour effectuer des recherches.

Les exemples du document et les écrans présentés sont issus d'applications gérées par le logiciel documentaire Basisplus (IDB).

# 2. Particularités des logiciels documentaires

## Information non structurée

Les logiciels documentaires gèrent essentiellement des informations non structurées: un texte en langage naturel, des images, ...

L'unité de recherche et de saisie est l' «**unité documentaire**» qui est composée :

- du **texte** en langage naturel ;
- de **l'image du document ;**
- des **informations descriptives** servant de critères de recherche pour retrouver le texte ou l'image du document (date, n° de document, auteur, type de document, ...).

Les bases documentaires gèrent donc des milliers d'unités documentaires indépendantes les unes des autres.
En fait, depuis le début des années 90, les concepteurs de ces logiciels proposent des possibilités de structuration de l'information. Ces possibilités viendraient s'ajouter autour du fichier contenant le texte.
Cependant, l'utilisation de ces possibilités est peu réaliste dans le cadre d'un projet documentaire reprenant des collections de textes ou d'images de documents. En effet, cette structuration complexifie tellement les recherches qu'on ne peut l'imposer aux utilisateurs de logiciels documentaires qui sont souvent des utilisateurs qui ne connaissent pas bien le contenu de la base.

## Volume

Le contenu de ces bases étant des **textes et/ou des images**, le volume des bases documentaires est souvent énorme.

Vu la diminution des coûts de la mémoire et l'augmentation de la capacité des mémoires, cet aspect est moins un frein au développement d'applications documentaires qu'il y a une dizaine d'années.

Certaines informations sont redondantes d'une unité documentaire à l'autre. Par exemple, si Dupond est l'auteur de 150 textes de la base, son nom se retrouvera 150 fois dans la base. Cependant, les redondances des informations descriptives représentent une proportion si minime par rapport au volume de la base qu'elle n'est plus prise en considération.

## Zone de longueur indéfinie et pratiquement illimitée

Le contenu des bases étant des documents sous forme de textes ou d'images, il est naturellement impossible de connaître au préalable les longueurs de ces documents. Certains seront très courts, d'autres pourront atteindre des dizaines de pages.

Les logiciels documentaires conçus pour gérer des textes de longueur variable permettent de définir des **zones de longueur variable**, c'est-à-dire qui occupent une place mémoire variable en fonction des données introduites.
Quand la zone reste vide pour un document, la taille de la zone est réduite au minimum. Quand le texte contient 10 pages, le volume occupé est celui d'un texte de 10 pages.

Dans le cas de Basisplus, la limite est de 128.000.000 caractères pour 1 texte, ce qui représente plusieurs milliers de pages.

Dans la pratique, il est tout à fait déconseillé de stocker dans une base des textes aussi longs. Aucun utilisateur ne va lire des centaines de pages à l'écran. L'idéal est d'avoir des unités documentaires de 2 ou 3 pages et de diviser les longs textes en unités documentaires plus fines.

## Zone multi-value

Une zone multi-value accepte plusieurs informations de même type différenciées par un **séparateur**. Par exemple, quand un document concerne plusieurs employeurs, les matricules de ces employeurs sont repris dans une zone MATRICULE et séparés par un séparateur (dans l'exemple, la virgule).

Ex. : 1.265.266-48,598.456-52,895.155-36

Cette fonctionnalité est indispensable puisque la structure de la base est en général mono-fichier. Elle prend le contrepied des principes de normalisation appliqués dans les bases de données relationnelles.

Cet avantage peut être considéré comme primordial dans certaines applications. Il est déjà arrivé que des applications ne contenant aucun texte soient développées à l'aide de logiciels documentaires uniquement pour profiter de cette particularité.

Encore une fois, cette possibilité rend les recherches plus simples pour l'utilisateur qui interroge toujours le même fichier.

## Unité documentaire et identifiant logique

Le choix de l'unité documentaire dépend de la longueur des documents, de la manipulation que l'utilisateur souhaite obtenir du document et du résultat souhaité.

Idéalement, l'unité documentaire ne peut dépasser 2 ou 3 pages. En effet, l'utilisateur préfère imprimer le document s'il est trop long.

Le document de 15 pages est donc découpé en plusieurs unités documentaires différentes. Par exemple, dans le cas des procès-verbaux qui ont 15 pages ou plus, l'unité choisie est 1 point du procès-verbal. Les points traitant de sujets divers, cette structure correspond aussi mieux aux besoins de recherche. Dans le cadre de ce techno, l'unité serait le paragraphe.

En ce qui concerne les textes de loi, l'unité documentaire peut être l'unité de parution au Moniteur s'il s'agit simplement de les consulter mais l'unité doit être plus fine s'il s'agit d'établir des liens entre la partie du texte qui a été modifiée et le texte modificateur - ou d'établir des liens entre les différentes versions ou coordinations d'un même texte.

Quand l'unité documentaire a été choisie pour chacun des types de documents à reprendre dans la base, il faut déterminer un identifiant logique unique.

Cet identifiant logique permet de vérifier lors de la saisie des documents si le document a déjà été saisi ou pas. Plus la base contient de types de documents différents, plus il est difficile de trouver un identifiant, commun à l'ensemble des unités de la base. Cet identifiant est toujours une concaténation d'informations descriptives. Il reprend en général le type du document + la date du document + le numéro du document (s'il existe, ce qui est rare) + d'autres informations variables suivant le contenu de la base comme le sous-type, la législature et la session parlementaire, le numéro de l'article de loi et il faut même parfois ajouter un nom d'auteur ou de revue.

Quand les types de documents sont nombreux, il est très difficile de trouver un identifiant commun et l'identifiant est composé d'une partie commune et d'une partie qui dépend du type de documents.

## Connexion avec d'autres logiciels

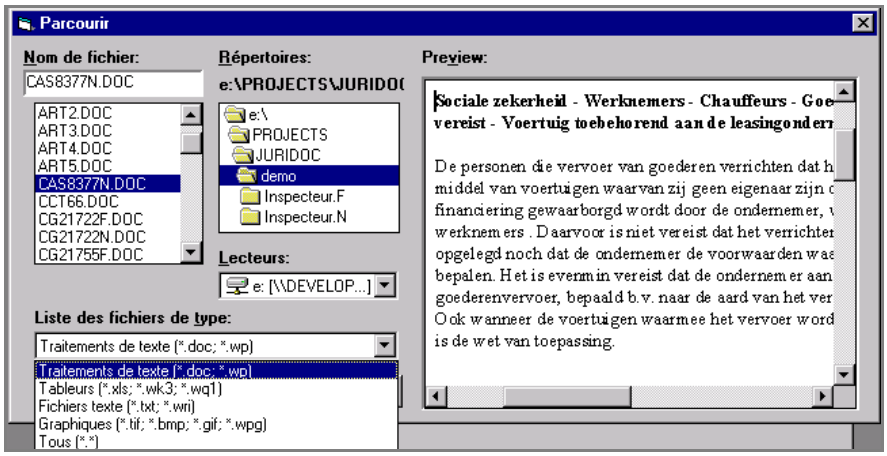
Les logiciels doivent proposer une intégration facile avec les outils bureautiques du marché: **traitements de texte, logiciels de scanning, tableurs, OCR**, …

En effet, il doit être simple d'introduire des documents de traitement de texte, des images, des tableaux, ... dans la base. Les opérations se passent dans l'ordre suivant :

- les personnes rédigent leur document à l'aide de leur logiciel habituel (traitement de texte, tableur, ...),
- si nécessaire, l'image scannée peut être «lue» par un logiciel OCR. L'OCR (Optical character recognition) reconnaît les caractères de l'image et transforme l'image d'un document en «texte» manipulable par un traitement de texte.

Le responsable de la base crée l'identifiant du document et y «attache» le texte, le tableau ou l'image.

Exemple d'écran issu du logiciel Basisplus : cet écran permet d'attacher un document traitement de texte, un tableau ou une image à une unité documentaire :




Quand un document est «attaché», il est chargé dans la base et une indexation automatique reprend tous les termes du texte dans un index.

## Paramétrage

Certaines fonctionnalités peuvent être simplement paramétrées.

### *Détermination des mots vides*

Un mot vide est un mot vide de sens (le, la, un, une, de, ...) ou, plus précisément, un mot repris dans tous les documents et ne pouvant servir de critère de recherche. Les mots vides ne sont pas indexés. Pour rappel, tous les termes des textes sont indexés, tous sauf ceux faisant partie de la liste des mots vides. Cela permet d'alléger sensiblement la base des index.

Le Data Base Administrator peut définir lui-même la liste des mots vides liés à une base. Cela permet d'affiner la liste en tenant compte des termes de la base. Par exemple, dans une base ne reprenant que des questions parlementaires, le terme "question" doit devenir un mot vide car il apparaît dans tous les documents et ne peut plus être considéré comme critère de recherche.

Dans une base de données multilingue, cette liste accumule les mots vides des différentes langues. Les termes qui sont vides dans une langue et significatifs dans une autre ne peuvent pas être repris dans cette liste.

### *Détermination des séparateurs*

Un séparateur est un caractère permettant de séparer les différents termes d'index et de déterminer les entrées de l'index. Le DBA peut également définir lui-même la liste des séparateurs liés à une base. Le "blanc" doit nécessairement être défini comme séparateur. Par contre, le point n'est pas un séparateur nécessairement intéressant. En effet, il n'est pas séparateur quand il sépare des majuscules formant un sigle comme dans O.N.S.S.

### *L'affichage de la liste des documents résultants*

L'utilisateur doit pouvoir facilement déterminer le tri des documents dans la liste de documents résultants. Il choisit un tri sur l'une des zones (ascendant ou descendant) ou un tri par ordre de pertinence. La pertinence est déterminée par le nombre d'occurrences du terme recherché dans un document.

### *Impression des documents*

L'utilisateur peut choisir l'impression d'un document ou de l'ensemble des documents répondant à une question. Il peut demander l'impression des zones descriptives d'un document ou des unités documentaires complètes, c'est-à-dire les zones descriptives + la zone texte (ou image du document).

### *Recherche dans les versions des textes*

L'utilisateur choisit d'effectuer ses recherches sur les dernières versions des textes ou dans toute la base, c'est-à-dire dans toutes les versions des textes.

## 3. Recherche

Les logiciels documentaires sont conçus pour la recherche d'informations. Ils permettent un éventail de types de recherches différentes qui ne doivent pas être préalablement définies.

En plus des critères de recherche classiques, l'utilisateur a, à sa disposition, des critères de recherche linguistiques basés sur la forme du mot, sur la situation du mot dans la phrase ou sur le contenu du texte.

Avant d'approfondir ces possibilités, il convient de souligner la différence entre une recherche dans une base de données structurée et une recherche dans une base textuelle où l'information est en grande partie non structurée. Quand un utilisateur fait une recherche dans une base structurée, sa recherche est prévue et il est sûr d'obtenir une réponse exacte, qu'elle soit significative ou nulle. En effet, si la réponse est nulle, l'utilisateur peut être sûr que le cas recherché ne se trouve pas dans la base.

Quand un utilisateur fait une recherche dans une base textuelle, il peut rarement être certain que la réponse ne contient pas de «silences».

## Silences et bruits

Les **silences** sont les documents qui devraient répondre à la question posée mais qui n'apparaissent pas dans la liste de documents répondant à la question.

Ex. : L'utilisateur demande des documents relatifs à des exonérations pour cas de force majeure.

Il n'obtient pas les documents traitant d'exonération pour cause d'incendie.

Les silences sont inévitables : ils proviennent en général du fait que l'utilisateur ne connaît pas le contenu de la base.

Du côté utilisateur, il est toujours préférable de poser sa question de plusieurs manières différentes et nous verrons plus tard qu'il y a toujours au moins deux façons possibles de poser une question.

Le but est d'éviter au maximum les silences en définissant des zones descriptives reprenant des informations précises du document (date, nom, ...) et en prévoyant plusieurs zones à valeur ajoutée par rapport au texte: par exemple, thesaurus, mots clés et résumé.

Les **bruits** sont des documents qui répondent à une question alors qu'ils ne devraient pas.

Ex. : L'utilisateur s'intéresse au trafic urbain et demande des documents relatifs aux embouteillages.

Il obtient entre autres des documents parlant de la mise en bouteille.

Les bruits peuvent également devenir gênants et découragent l'utilisateur avec des documents qui ne l'intéressent pas. Les logiciels documentaires qui font systématiquement des recherches approximatives de termes sont parfois pénibles au niveau bruit. Certains logiciels ajoutent à la recherche de l'utilisateur tous les termes qui ressemblent au terme recherché à une lettre près.

Lors de la mise en oeuvre d'une base documentaire, le but est de trouver la meilleure adéquation pour éviter au maximum les silences sans provoquer pour autant trop de bruit.

## Une recherche, un dialogue

Comme nous venons de le voir, nous ne pouvons considérer qu'une consultation = une réponse.

L'utilisateur doit chercher de plusieurs manières différentes et/ou faire preuve d'imagination pour poser sa question différemment.

La moyenne de 2,5 questions pour une réponse est tout à fait honorable.

Il arrive aussi fréquemment qu'après une première recherche, l'utilisateur obtienne trop de documents répondant ou qu'aucun document ne réponde à sa question. Il devra alors affiner ou élargir sa recherche.

## Mode de recherche habituel

L'utilisateur a à sa disposition un écran de recherche qui lui permet d'effectuer des recherches sur une ou plusieurs zones de façon combinée. Il peut également demander plusieurs critères coexistant dans une même zone.

Ex. : Durant et Dupont dans la zone NOM et 20.01.2000 dans la zone DATE

Les opérateurs de recherche :

- les **opérateurs arithmétiques**: multiplication (* ), division (/), addition (+), soustraction (-), exponentiation (**)
- les **opérateurs relationnels** : >, <, entre
- les **opérateurs booléens**: AND, OR, AND NOT
- **fonctions statistiques**: AVG, MIN, MAX, SUM, COUNT

## Phrases ou suite de caractères

L'utilisateur peut effectuer des recherches sur des phrases ou parties de phrases

Ex. : Loi sur les pensions
Ex. : 90%, 2nd, …

## Troncature à gauche, à droite

L'utilisateur recherche un terme dans le texte.

### *Troncature à droite*

Ex. : pension*

Il obtient les documents contenant le terme pension au singulier et au pluriel, le terme pensionné, ...

La troncature à droite est indispensable car l'utilisateur ne sait jamais si le terme qu'il recherche est au singulier ou au pluriel dans le texte.

### *Troncature à gauche*

Ex. : *vergoeding

Il obtient les documents contenant les termes vergoeding, vooropzegvergoeding, arbeidsongevallenvergoeding, autokilometervergoeding, begrafenisonkostenvergoeding, beroepsvergoeding, beschermingsvergoeding, fietskilometervergoeding, huisvestings-vergoeding, inschakelingsvergoeding, integratievergoeding, invaliditeitsvergoeding, kampvergoeding, kantoorvergoeding, kilometervergoeding, …

La troncature à gauche est indispensable dans les applications qui reprennent des documents en néerlandais ou en allemand.

## Thesaurus

### *Qu'est-ce qu'un thesaurus ?*

Partons de la notion bien connue du «dictionnaire».

Un dictionnaire est un ensemble de termes «isolés» qui sont expliqués. La présentation est un ordre alphabétique de termes.

Par contre, un thésaurus est un ensemble de termes reliés entre eux par des liens explicitement définis.

Un thésaurus est donc un dictionnaire dans lequel chaque terme est au moins relié à un autre.

Ces liens peuvent être :

- verticaux: les termes sont hiérarchisés : un terme générique est relié à ses spécifiques qui peuvent eux-mêmes avoir des spécifiques… Le nombre de niveaux de spécificité n'est en théorie pas limité. En pratique, l'utilisateur risque d'être perdu dans l'arborescence si le thésaurus contient plus de 10 niveaux ;
- horizontaux: un lien horizontal est par exemple la synonymie entre 2 termes.

Le nombre de termes d'un thésaurus n'est pas non plus limité. En pratique, quand un utilisateur doit se servir d'un thésaurus trop volumineux (certains thésaurus ont 10.000 – 12.000 termes), il se limite souvent à une petite partie du thésaurus.

### *Quelques types de liens fréquemment utilisés*

- Généricité et spécificité

Ex : Secteur d'activité
- Commerce
  - Carrosserie
  - Garage
  - Commerce de détail
  - …
- Education

Enseignement
Enseignement primaire
Enseignement supérieur
…
Formation permanente
Auto-école
…
…
…

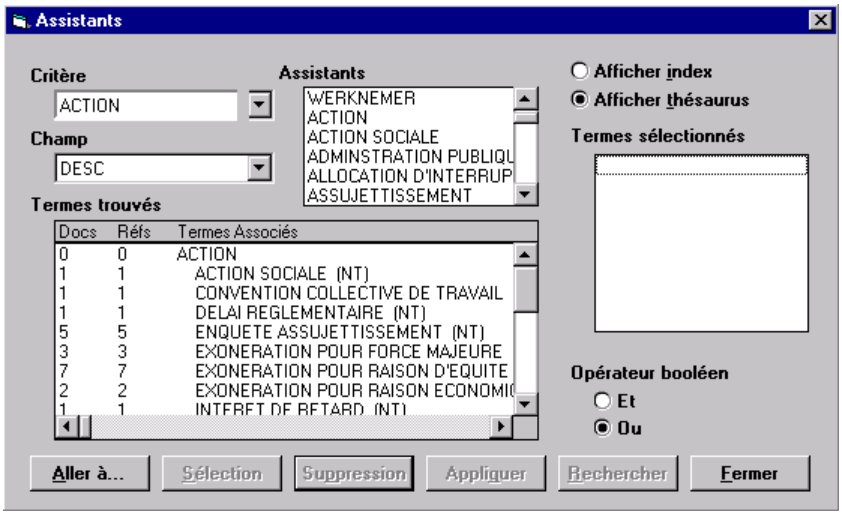
- Synonymie
  ex. : nettoyeuse - personnel de nettoyage
- Corrélation
  2 termes sont corrélats quand il est opportun de conseiller de consulter l'autre terme également.
  Ex. : Recherche scientifique
  voir aussi : Engagement dans la recherche scientifique
- Abréviation
  Ex. : Office National de Sécurité Sociale - ONSS
- Définition
  Ex. : garde dormante: garde autorisée à dormir sur le lieu du travail
- Historique
  Lie un ancien terme à son équivalent plus récent
  Ex. : sourd – mal entendant
- Synonymie translinguale
  Ex. : enseignement = onderwijs

## *Consultation d'une base à l'aide d'un thesaurus*

Un utilisateur peut consulter une base en demandant un terme du thésaurus, que ce terme soit défini comme générique, spécifique, synonyme, abréviation, terme historique, …

Un premier avantage revient à l'adjonction dans son équation de recherche :

- des synonymes,
- des spécifiques du terme recherché.



Par exemple, un utilisateur qui lance une recherche en introduisant “action” comme critère de recherche, trouvera

- les documents indexés par ce terme (dans l’écran ci-dessus, il n’y en a pas),

- les documents indexés par un spécifique de ce terme.

Dans l’exemple: action sociale, convention collective de travail, délai réglementaire, enquête assujettissement, exonération pour force majeure, exonération pour raison d’équité, exonération pour raison économique, intérêt de retard, ...

Un deuxième avantage est l’adjonction à son équation de recherche

de la traduction du terme;

des synonymes de la traduction;

des spécifiques des termes traduits.

Dans l’exemple supra:

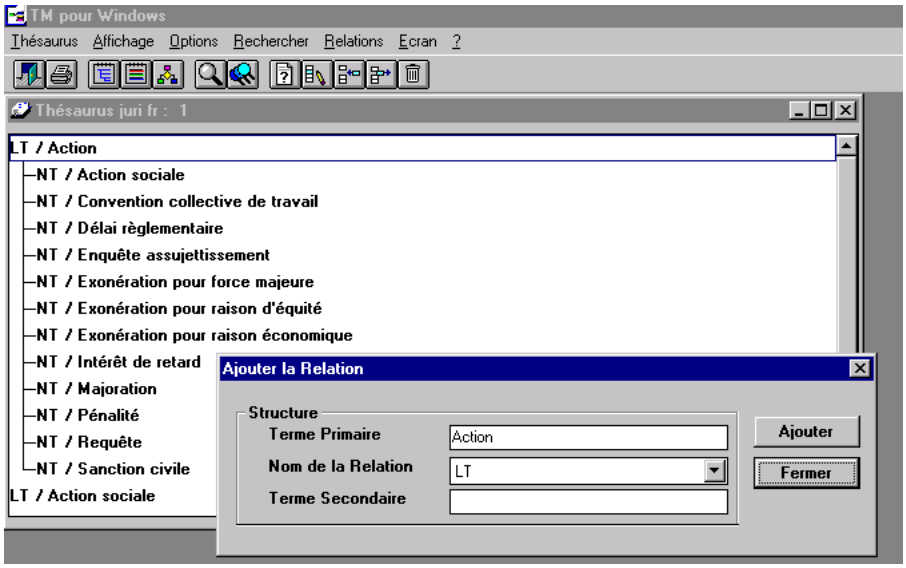
- les documents indexés par la traduction du terme: actie

- les documents indexés par le terme aktie, défini comme synonyme

- les documents indexés par un spécifique du terme traduit: sociale actie, collectieve arbeidsovereenkomst, reglementaire vervaldag, onderwerpingsonderzoek, vrijstelling voor overmacht, vrijstelling voor billijkheidsredenen, vrijstelling voor economische redenen, verwijlinterest, ...

- il obtiendra donc des documents écrits en français et en néerlandais car on peut supposer qu’un utilisateur intéressé par un sujet est intéressé par des documents écrits dans les deux langues. Cette manière de procéder est surtout intéressante dans le cas des documents non traduits.

## *Gestion du thesaurus*




Un gestionnaire de thésaurus (dans l’exemple, Basis Tm) permet d’ajouter et de supprimer un terme dans un thésaurus. Ce module permet de modifier le thésaurus sans devoir le recharger entièrement.

## Une liste de zones établies par le DBA

- Liste de zones

Dans certains cas, il est préférable d'établir des listes de zones car l'utilisateur ne penserait pas à chercher lui-même dans les différentes zones.

De plus, cela évite à l'utilisateur de répéter plusieurs fois le terme ou la locution recherchée dans plusieurs zones de recherche.

Ex. : L'utilisateur peut souhaiter faire des recherches simultanément

dans la zone thesaurus
dans la zone texte
dans la zone intitulé

Les champs de cette liste doivent naturellement être tous du même type (date, alphabétique, numérique).

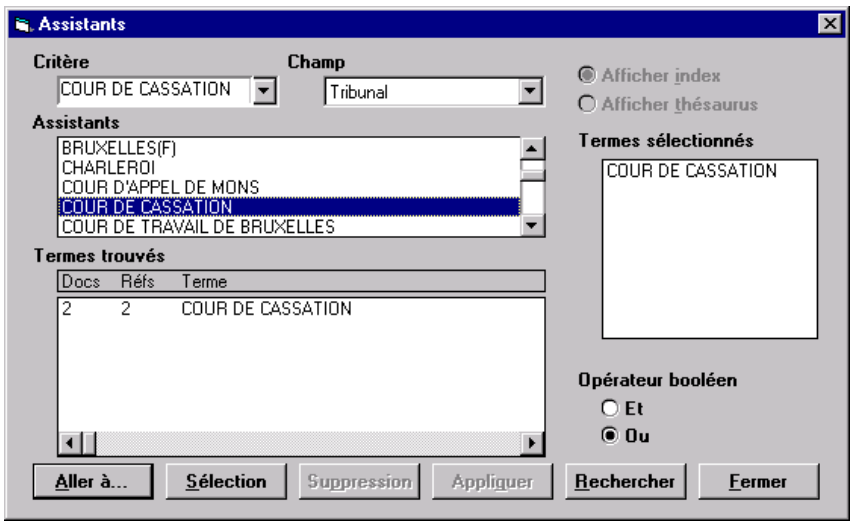
- une partie d'un champ

Certains champs peuvent être composés de plusieurs sous-parties, et l'utilisateur ne souhaite pas nécessairement faire une recherche sur l'ensemble du champ.

Ex. : un numéro de document: EXT/97.0192

## Recherche par les index : Recherche par assistants

Les assistants aident l'utilisateur à trouver ses critères de recherche en affichant le contenu des index par champ.




En choisissant un champ, l'utilisateur peut parcourir la liste des index établis à partir de ce champ. Il peut sélectionner un terme et obtenir le nombre de documents répondant à la question et le nombre de fois que le terme est repris dans l'ensemble des textes.

A partir de ces informations, il peut choisir ses critères de recherche et lancer la requête.

## Recherches secondaires : Lien entre les documents

Des liens entre les documents peuvent être établis à partir d'un document retrouvé.

### *Hypertexte*

L'utilisateur effectue une recherche et visualise 1 document. Dans le texte du document, il surligne un mot ou une suite de mots et lance une recherche.

La recherche porte sur tous les documents qui contiennent le mot ou la suite de mots surlignés. Ce type de recherche est intéressant notamment pour trouver :

- tous les textes faisant référence au même texte juridique,
- tous les documents portant un même titre de paragraphe ou sous-paragraphe (dans les instructions aux employeurs, l'unité documentaire est le sous-paragraphe, l'ensemble du paragraphe ou du chapitre peut être retrouvé de cette façon).

### *Version*

La notion de version peut être utilisée à plusieurs escients :

#### un même document évolue

Suivant le cas, le document initial n'a plus de valeur ou garde une valeur historique. Par exemple, dans une base juridique, les juristes souhaitent garder toutes les versions car ils doivent connaître le texte qui était en vigueur à une date du passé.

L'utilisateur peut avoir le choix entre effectuer sa recherche sur toutes les versions ou sur les dernières versions des documents.

#### version linguistique

La gestion des versions permet à l'utilisateur de passer facilement d'un document dans une langue au document dans l'autre langue. Les personnes bilingues préfèrent souvent lire le document dans sa langue source. Après une recherche, l'utilisateur se positionne dans la liste sur le document qui l'intéresse et peut demander d'afficher les versions du document.

#### dossier

La notion de version peut également être utilisée pour constituer des dossiers à partir des différentes unités documentaires reprenant les pièces du dossier. L'utilisateur peut retrouver ainsi toutes les pièces d'une affaire juridique ou tous les échanges de courrier relatifs au même sujet.

### *Lien entre valeur de champ identique*

Sur base d'un document trouvé, il y a moyen de trouver les documents dont une des zones est identique. L'utilisateur peut sélectionner le champ liant et il obtient la liste des documents pour lesquels le contenu de ce champ est identique au contenu du document de départ.

Exemple: Cette option est utilisée pour retrouver tous les documents ayant une source légale identique. Donc, cela permet d'obtenir tous les documents modifiant un certain article de la loi de base.

## 4. Intégration dans l'environnement Internet/Intranet

Internet a présenté une véritable opportunité pour les logiciels documentaires qui sont rapidement devenus compatibles avec les principaux navigateurs du marché. L'interface d'accès par défaut présente des écrans HTML simples, composés d'une barre d'icônes, permettant à l'utilisateur d'effectuer des recherches, de consulter des lots résultats et d'afficher les documents.

Des formulaires de recherche multi-critère, comprenant des opérateurs booléens et des critères de tri sont générés automatiquement. Via ces formulaires, l'utilisateur peut effectuer des recherches portant simultanément sur des zones descriptives et le contenu textuel des documents, avec toute la puissance du moteur d'un logiciel documentaire, incluant technologie de recherche texte intégral avec proximité, indexation automatique, recherche par concepts, gestion de bases de thesaurus.

## 5. Conclusions

En conclusion, les logiciels documentaires gèrent des **masses d'informations non structurées** et ont développé des outils de recherche aidant l'utilisateur à y retrouver une information.

Le public se servant de ces logiciels est large, diversifié et parfois peu averti quant à l'utilisation d'outils informatiques. La **convivialité** et l'utilisation intuitive doivent prévaloir lors de l'implantation des applications même si une solution simple est parfois difficile à trouver.

Les **outils de recherche**, très développés dans ces logiciels, permettent d'effectuer des recherches dans des textes en se basant sur les **termes du texte**, sur le **sujet du document** (thesaurus, mots clés), sur les **liens entre les documents** (locution commune, gestion des versions et des dossiers, valeur identique d'un champ).

Des aides sont proposées à l'utilisateur par le biais de la consultation des assistants: recherche dans les différents **index** et affichage des index et du thesaurus.

Le **thesaurus** aide l'utilisateur à retrouver des documents relatifs à un sujet, les documents pouvant être rédigés en différentes langues.

Par toutes ces caractéristiques, ces logiciels documentaires sont de plus en plus utilisés comme outils de recherche dans les bases **INTRANET/INTERNET** pour guider l'internaute.

Ces deux technologies sont de toute évidence faites pour être intégrées l'une à l'autre: les logiciels documentaires apportent leurs techniques de recherche dans le domaine des systèmes d'information et les technologies INTRANET/INTERNET permettent, d'une part, de réduire considérablement les coûts de diffusion des bases documentaires et, d'autre part, d'envisager une diffusion beaucoup plus large des informations.