Manuel Utilisateur KB Crawl

KB CRAWL SAS - 10 rue Lionel Terray - 92508 Rueil-Malmaison Cedex (France) Tel : 01 41 29 05 02 Fax: 01 57 67 04 59 www.kbcrawl.net

SOMMAIRE

Les termes marqués d'un astérisque (*) sont définis dans le glossaire en fin de manuel.

Table des illustrations

1 Introduction à KB Crawl

1.1 Principales fonctionnalités

- KB Crawl est un outil d'assistance à la recherche et à l'exploration d'informations sur Internet*. Habituellement, lorsqu'un internaute* veut avoir accès à une information, il ouvre son navigateur et visualise les pages Web* qui l'intéressent puis analyse leur contenu soit pour en prendre connaissance, soit pour détecter d'éventuels changements à l'intérieur de ce contenu.
C'est le processus que l'on appelle « Veille ».
Ainsi, le veilleur va périodiquement rechercher de nouvelles informations, télécharge* et analyse les mêmes pages et passe une grande part de son temps à surfer sur le Web.

- KB Crawl permet d'effectuer automatiquement ces tâches répétitives et de reproduire le processus de veille depuis l'exploration des sites contenant des informations pertinentes jusqu'à la détection de tout changement à l'intérieur des documents qu'il contient afin d'en alerter le veilleur.

- Ces alertes sont mises en évidence au niveau de l'interface de KB Crawl afin d'avertir immédiatement l'utilisateur. Elles peuvent également être diffusées par courriel à une liste de contacts ou à des groupes de contacts.

- KB Crawl propose une palette d'outils dédiés à l'analyse des informations acquises durant la phase d'exploration appelée « crawl* ».
A chaque fois que KB Crawl inspecte un site Internet, il stocke le contenu textuel des pages qu'il explore dans une base de données*. Ceci permet d'effectuer des recherches par mots-clés (requêtes) afin de localiser précisément l'information recherchée : quelles pages contiennent cette information et où se trouve cette information à l'intérieur de ces pages.

- Outre la fonctionnalité de moteur de recherche, le stockage de ces contenus dans une base de données relationnelle permet de : visualiser des sites sous forme arborescente, mettre en évidence des mots-clés recherchés et de nouveaux mots-clés apparus (dans un browser* intégré à l'interface), classer les sites observés par dossiers, etc.

- Le module d'archivage de KB Crawl permet de stocker les différentes versions d'une page Web analysée puis de les consulter pour les comparer entre elles et suivre l'évolution de l'information à l'intérieur de ces pages.

- Le moteur de KB Crawl permet d'accéder à des pages du « Web invisible*» et de les surveiller en enregistrant au préalable les données à envoyer aux différents formulaires* rencontrés.

1.2 Performances : quelques ordres de grandeur

Internet supporte aujourd'hui plusieurs milliards de pages Web. 86 % de ces pages ne sont pas mises à jour régulièrement. Un chargé de veille documentaire, concurrentielle ou technologique est par conséquent amené à visiter des sites d'information 6 fois sur 7 pour rien.

Surveiller 30, 40 ou 100 sites quotidiennement conduit donc à consacrer une partie importante de son temps à "surfer" sans résultat exploitable. Pourtant, l'information évolue et 14% des sites connaissent des modifications plus d'une fois par semaine.
KB Crawl télécharge, analyse (parsing*) puis stocke le contenu des pages Web dans sa base de données à un rythme impossible à atteindre manuellement.

Les performances de ce traitement varient en fonction de plusieurs facteurs :

- la qualité de la connexion Internet du poste qui utilise KB Crawl,
- la réactivité du serveur hébergeant le site,
- la réactivité du provider* (fournisseur d'accès Internet),
- la rapidité du processeur de l'ordinateur (fréquence de l'horloge),
- la taille des pages qui sont téléchargées.

L'algorithme interne de KB Crawl rend négligeable la durée de la phase d'analyse d'une page par rapport au temps de téléchargement. Le stockage dans la base de données est quant à lui quasi-immédiat.

Exemples de temps de traitement (pour un même serveur) ci-dessous :

- en abscisse : la taille de la page (en octets),
- en ordonnée : le nombre de secondes pour effectuer le traitement.

Processeur : Pentium 800Mhz
Connexion : ADSL*, vitesse de téléchargement de 1024 Kbits/s (maximum)

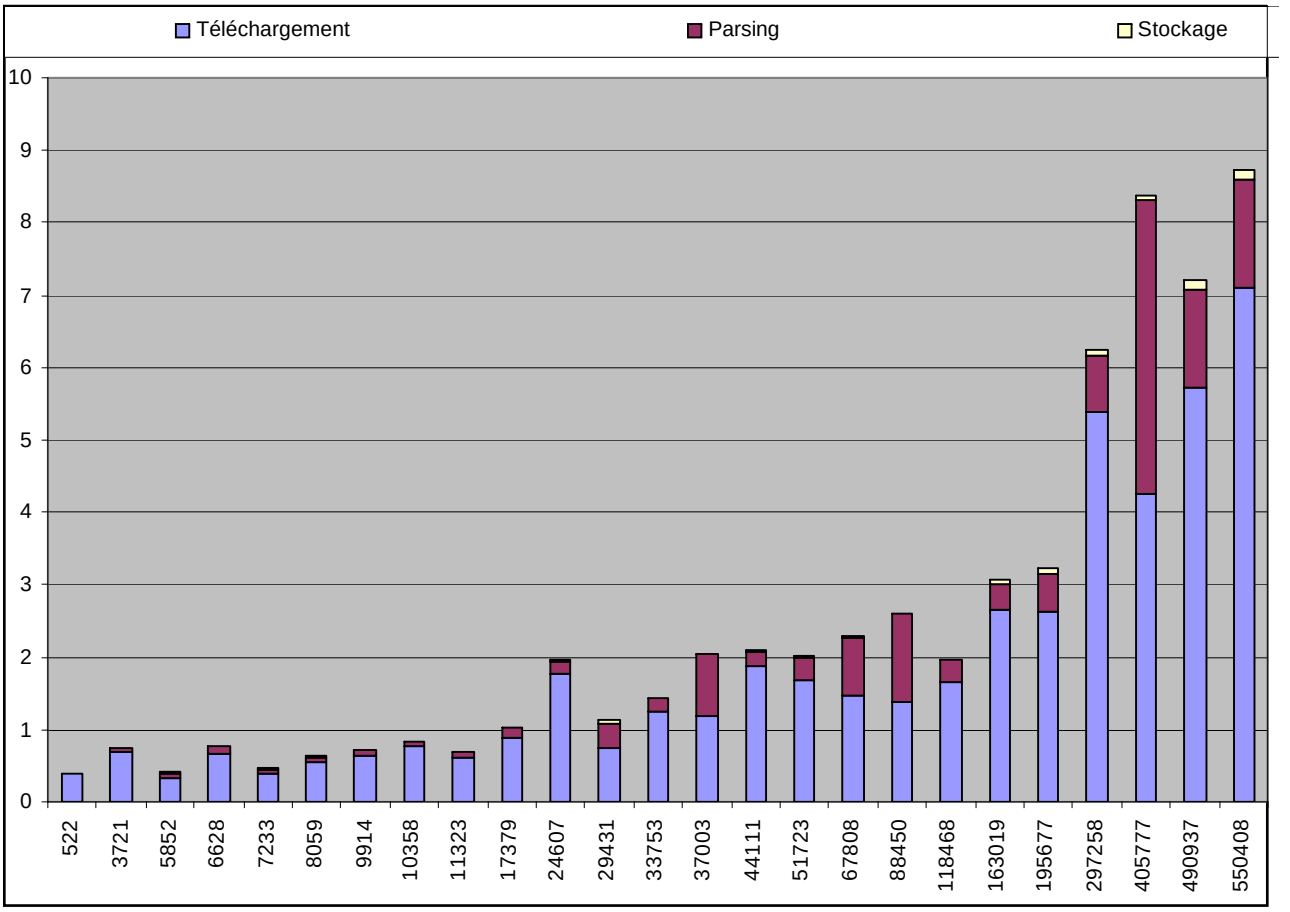


Figure 1 : Durée du crawl en fonction de la taille de la page.

KB Crawl passe plus de temps en moyenne à télécharger les pages qu'à les analyser.
Le rapport entre les deux est fonction des fluctuations du réseau.
Le temps de stockage est toujours négligeable.
Si on ne peut pas annoncer une durée fixe de traitement pour une page, puisque cette durée dépend de plusieurs facteurs : taille du fichier, réseau, réactivité du serveur Web, etc., on peut, avec l'expérience, retenir une moyenne de 1minute pour le crawl d'une centaine de pages.

1.3 Pré-requis matériel

Environnement : Microsoft Windows 2000, XP, 2003 Server, VISTA

Processeur : Pentium 1.5 Ghz ou processeur équivalent
Mémoire vive : 512 Mo (minimum), 1 Go (recommandé)
Connexion internet : ADSL conseillé)

1.4 Téléchargement

Les pages HTML* que l'on télécharge sont de très longues chaînes de caractères (suites d'octets) qui constituent le code HTML.
Le téléchargement est la première phase du traitement appelé « crawl* ».
Concrètement, KB Crawl envoie une requête HTTP à un serveur Web.
Une requête HTTP est une chaîne de caractères qui contient entre autres :

- l'adresse de l'hôte (serveur qui héberge les pages Web)
exemple : 217.147.236.241 (= adresse IP)
ou : www.kbcrawl.com (= alias qui redirige vers une adresse IP),

- le chemin complet du fichier que l'on souhaite télécharger
exemple : /products/KBCRAWL.htm
On peut aussi appeler /products parce que le serveur a une page définie par défaut pour ce répertoire.

Il arrive fréquemment qu'un téléchargement échoue, pour des raisons diverses :

- la page ne se situe pas à l'emplacement désigné (HTTP/1.1 404 Objet introuvable),
- un problème est survenu avec la connexion Internet (Socket* Error # 10061 Connection refused),
- le serveur qui doit renvoyer la page a eu un problème technique (HTTP/1.1 500 Erreur serveur interne).

Lorsqu'une erreur se produit durant le téléchargement, elle est immédiatement inscrite au journal de KB Crawl (cf. § 11). Il est parfois utile de consulter ce journal pour prendre connaissance du compte-rendu complet d'un crawl et voir si des difficultés particulières ont été rencontrées.

```
Socket Error # 0 No Error
Socket Error # 10004 Interrupted system call
Socket Error # 10009 Bad file number
Socket Error # 10013 Permission denied
Socket Error # 10014 Bad address
Socket Error # 10022 Invalid argument
Socket Error # 10024 Too many open files
Socket Error # 10035 Operation would block
Socket Error # 10036 Operation now in progress
Socket Error # 10037 Operation already in progress
Socket Error # 10038 Socket operation on non-socket
Socket Error # 10039 Destination address required
Socket Error # 10040 Message too long
Socket Error # 10041 Protocol wrong type for socket
Socket Error # 10042 Bad protocol option
Socket Error # 10043 Protocol not supported
Socket Error # 10044 Socket type not supported
Socket Error # 10045 Operation not supported on socket
Socket Error # 10046 Protocol family not supported
Socket Error # 10047 Address family not supported by protocol family
```

Socket Error # 10048 Address already in use
Socket Error # 10049 Can't assign requested address
Socket Error # 10050 Network is down
Socket Error # 10051 Network is unreachable
Socket Error # 10052 Net dropped connection or reset
Socket Error # 10053 Software caused connection abort
Socket Error # 10054 Connection reset by peer
Socket Error # 10055 No buffer space available
Socket Error # 10056 Socket is already connected
Socket Error # 10057 Socket is not connected
Socket Error # 10058 Can't send after socket shutdown
Socket Error # 10059 Too many references, can't splice
Socket Error # 10060 Connection timed out
Socket Error # 10061 Connection refused
Socket Error # 10062 Too many levels of symbolic links
Socket Error # 10063 File name too long
Socket Error # 10064 Host is down
Socket Error # 10065 No Route to Host
Socket Error # 10066 Directory not empty
Socket Error # 10067 Too many processes
Socket Error # 10068 Too many users
Socket Error # 10069 Disc Quota Exceeded
Socket Error # 10070 Stale NFS file handle
Socket Error # 10091 Network SubSystem is unavailable
Socket Error # 10092 WINSOCK DLL Version out of range
Socket Error # 10093 Successful WSASTARTUP not yet performed
Socket Error # 10071 Too many levels of remote in path
Socket Error # 11001 Host not found
Socket Error # 11002 Non-Authoritative Host not found
Socket Error # 11003 Non-Recoverable errors: FORMERR, REFUSED, NOTIMP
Socket Error # 11004* Valid name, no data record of requested type
Socket Error # 11004* No address, look for MX record

Tableau 1 : Liste des « socket errors » que l'on peut rencontrer lors d'un téléchargement

L'erreur la plus fréquemment rencontrée est « HTTP/1.1 404 Objet introuvable » qui n'est pas une « socket error » mais une erreur signalée dans l'entête de la réponse HTTP* du serveur qui avertit qu'il n'a pu trouver la page.

Remarque : Si une page contient « n » liens vers d'autres pages et qu'une ou plusieurs de ces pages n'ont pu être téléchargées à cause d'une erreur, cela ne change rien au fait que cette page contient « n » liens. Un lien déclaré comme présent dans une page ne veut pas dire que ce lien est valide. Par contre, si une page n'a pu être téléchargée, KB Crawl nous alertera en la faisant apparaître comme supprimée.

1.5 Parsing*

Les pages Web telles qu'on les visualise dans un navigateur sont issues de la traduction de leur code HTML*.
KB Crawl parcourt systématiquement l'intégralité de ce code pour l'analyser, c'est ce que l'on appelle le « parsing ».

1.5.1 Quelques notions de HTML

Le code HTML d'une page Web est constitué de balises qui servent à déterminer où commence et où finit une instruction du langage.
Exemple :
<B>Bonjour</B>
<B> est une balise qui signifie caractère gras (Bold)
</B> ferme cette balise
Résultat : tous les caractères placés à l'intérieur de la balise sont en gras.
Dans un browser, la traduction de ce code donnera :
Bonjour

A l'intérieur d'une page, certaines balises contiennent des informations qui sont pertinentes au regard de la recherche, d'autres non.

Quelques exemples :

<body bgcolor="#990000">

Cette balise <body > comporte un attribut (bgcolor) qui donne la valeur de la couleur de fond du corps de la page, elle ne présente pas d'intérêt particulier pour KB Crawl.

<meta name="description" content="Vous êtes à la recherche d'un logiciel de veille, venez découvrir le site KB Crawl, votre spécialiste de la surveillance des sites Internet.">

La balise meta contient deux types de contenus : « name » étant une propriété, en l'occurrence il s'agit de la meta balise de description du contenu de la page, et « content » qui est la valeur de la propriété (ici la description de la page sous forme de courte synthèse textuelle).

Les informations contenues dans ce « content »sont porteuses de sens, car elles constituent la description de la page qui contient des mots-clés à stocker.

```
<a href="profile.html" target="_blank">
```

La balise <a> est également intéressante car elle contient un lien (« profile.html ») qui va éventuellement servir à explorer la suite du site.
Remarque : Dans le corps de la page, les caractères qui se trouvent en dehors des balises constituent le texte que l'on voit apparaître :

```
<td valign="top" colspan="2"> <p><font face="Verdana, Arial, Helvetica, sans-serif" size="3"
color="#FFCC99"><b><a href="products/index.htm" target="_blank">
<br>Produits KB Crawl </a></b></font> <br><br></p></td>
```

Les mots du texte qui seront visibles dans la page apparaissant ici en rouge. Ces mots seront récupérés lors du parsing* puis stockés dans la base de données de KB Crawl.
Remarque : Ce texte apparaîtra sous forme d'un lien qui renverra sur la page http://www.kbcrawl.com/products/index.htm
Ce lien sera également récupéré.

1.5.2 Grammaire HTML

Quand KB Crawl effectue un parsing, il parcourt entièrement le code HTML de la page qu'il vient de télécharger et scrute chaque balise.
Certaines sont ignorées car sans intérêt particulier et d'autres sont analysées en détail pour extraire les informations qu'elles contiennent.
Pour faire cette distinction, KB Crawl possède sa propre grammaire HTML, c'est à dire une liste de balises dont il sait qu'elles contiennent des données qu'il doit analyser pour en extraire les informations puis les stocker dans sa base de données.

KB Crawl est livré avec une grammaire par défaut.

Balise ouvrante	Nom de variable ou de fonction	Balise fermante
<A	TARGET	>
<A	HREF	>
<A	WINDOW.OPEN	>
<APPLET	"URL" VALUE	</APPLET>
<AREA	TARGET	>
<AREA	HREF	>
<BASE	HREF	>
<BODY	BACKGROUND	>
<EMBED	SRC	>
<FORM	ACTION	>
<FRAME	SRC	>
<IMG	SRC	>
<LINK	HREF	>
<META	URL	>
<META	DESCRIPTION	>
<META	KEYWORDS	>
<OPTION	RECUPALL	>
<OPTION	VALUE	>

<PARAM	VALUE	>
<SCRIPT	FUNCTION	</SCRIPT>
<SCRIPT	ACTION	</SCRIPT>
<SCRIPT	HREF	</SCRIPT>
<SCRIPT	SRC	</SCRIPT>
<SCRIPT	MAKEMENU	</SCRIPT>
<SCRIPT	WINDOW.OPEN	</SCRIPT>
<STYLE	XXX	</STYLE>
<TITLE	RECUPALL	</TITLE>
IFRAME	SRC	</IFRAME>
ILAYER	SRC	</ILAYER>
{	TARGET	}
{	ACTION	}
{	LOCATION.HREF	}
{	HREF	}
{	SRC	}
{	URL	}
{	ACTION	}
{	LOCATION.REPLACE	}
{	OPEN	}
{	OPENPOPUP	}

Tableau 2 : Grammaire par défaut de KB Crawl

1.6 Récursivité et profondeur

Le processus de parsing a deux vocations :

- extraire tous les mots visibles et non visibles de la page,
- extraire tous les liens vers d'autres pages.

Un crawl a toujours un point de départ : une adresse Internet à partir de laquelle KB Crawl débute son exploration. Cette page est analysée selon le processus décrit plus haut puis tous les liens de cette page sont stockés temporairement.
Chaque lien faisant partie de cette collection est unique et constitue un nouveau point de départ pour KB Crawl qui peut ainsi reproduire le même processus (téléchargement, parsing, stockage) pour chacun d'eux, et ainsi de suite. C'est ce qu'on appelle un processus récursif.

Il est important de définir une limite pour cette exploration et ceci pour plusieurs raisons :

- seul un ensemble bien délimité de pages est en général intéressant (quelques fois même, une seule page Internet fait l'objet de la veille ou surveillance),
- le temps de l'exploration dépend du nombre de pages,
- tout stockage représente un coût en termes de place sur le disque dur,
- les liens d'un site Internet peuvent conduire vers un autre site et ainsi de suite, ce qui pourrait amener à « aspirer » tout le Web !

On considère deux types de liens relatifs à un même site Internet :

- les liens internes : ils permettent de naviguer vers des pages du même site,
- les liens externes : ils permettent de naviguer vers des pages d'un autre site.

Deux URL sont dites « du même site » si elles ont le même nom de domaine* ou si le domaine de l'une est le sous domaine de l'autre.

Exemple :
http://www.kbcrawl.com/KBCrawl/index.php et
http://www.kbcrawl.com/actualité.html
appartiennent au même site parce que leur nom de domaine « www.kbcrawl.com » est le même.

De même :
http://www.kbcrawl.com/KBCrawl/index.php et
http://www.mail.kbcrawl.com/KBCrawl/index.php
appartiennent au même site parce que www.mail.kbcrawl.com est un sous domaine de « www.kbcrawl.com ».

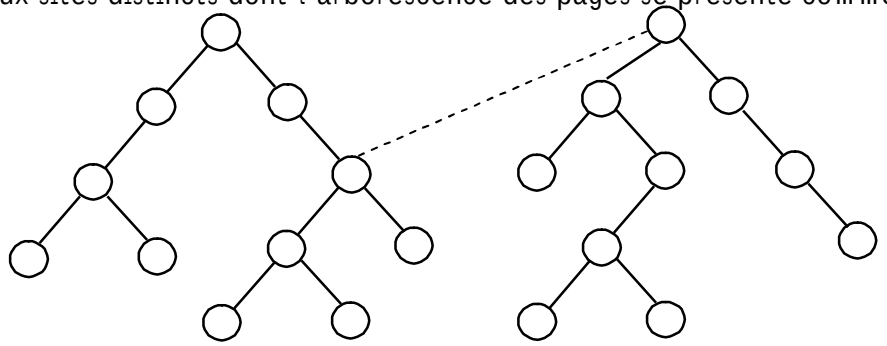
Il existe trois différents types de profondeur :

- La profondeur de page : c'est le nombre maximum de niveaux parcourus à l'intérieur du site de départ.
- La profondeur de site : c'est le nombre maximum de sites différents qui peuvent être explorés.
- La profondeur de page depuis les liens externes : c'est la profondeur de page qui s'applique dès lors que l'on explore un site différent du site de départ.

Par défaut, les valeurs proposées par KB crawl sont 0 pour ces trois paramètres.
A chaque fois que KB Crawl explore une page fille, il incrémente s'il y a lieu les compteurs de niveaux parcourus (qui valent 0 sur la page de départ), ce qui lui permet de s'arrêter quand ces compteurs atteignent la limite définie par l'utilisateur.
Il est possible également de fixer une limite au nombre total de pages explorées.

Exemples :

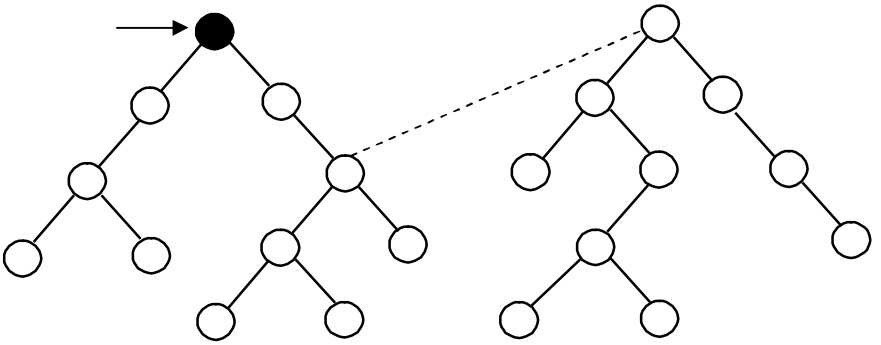
Soient deux sites distincts dont l'arborescence des pages se présente comme ceci :



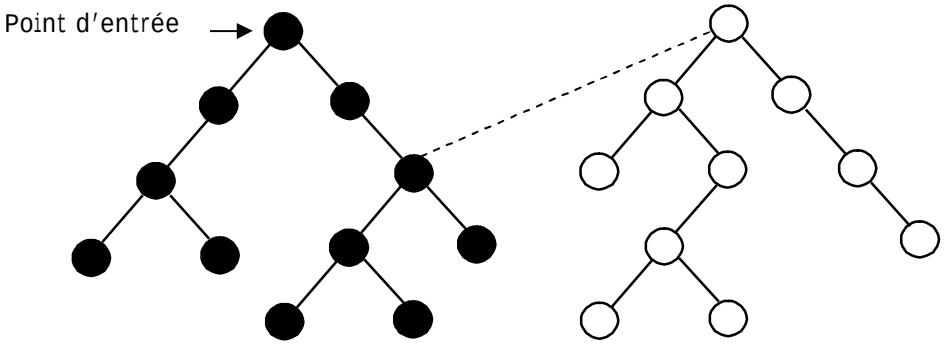
Dans les cas suivants, on colorie en noir les pages qui seront explorées.
Remarque : Le lien en pointillés mène du premier site vers la page d'accueil du second.

Cas 1 :
Profondeur de page : 0
Profondeur de site : 0
Profondeur de page depuis les liens externes : 0

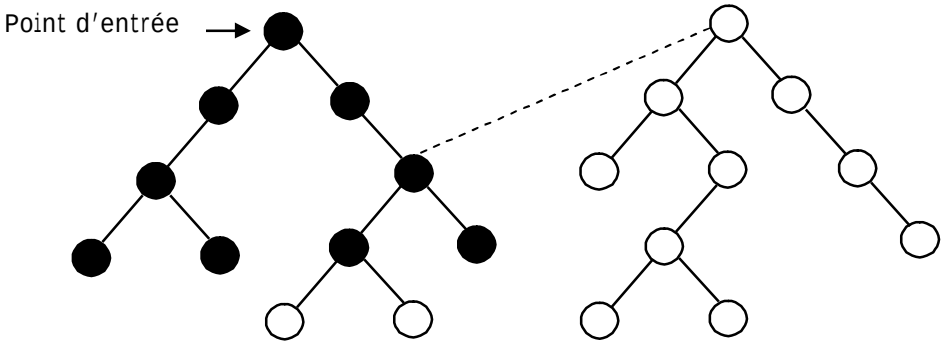
Point d'entrée



Cas 2 :
Profondeur de page : Infinie
Profondeur de site : 0
Profondeur de page depuis les liens externes : 0

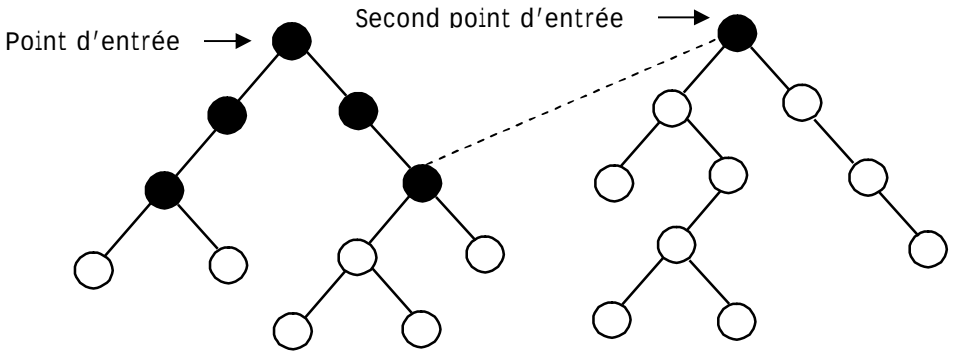


Cas 3 :
Profondeur de page : 3
Profondeur de site : 0
Profondeur de page depuis les liens externes : 0

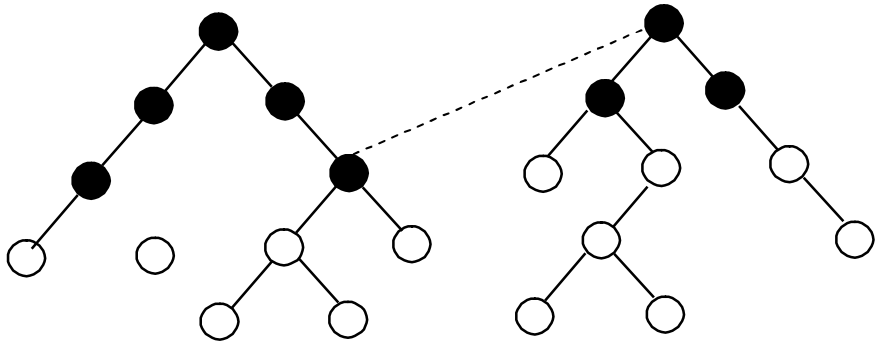


Cas 4 :
Profondeur de page : 2

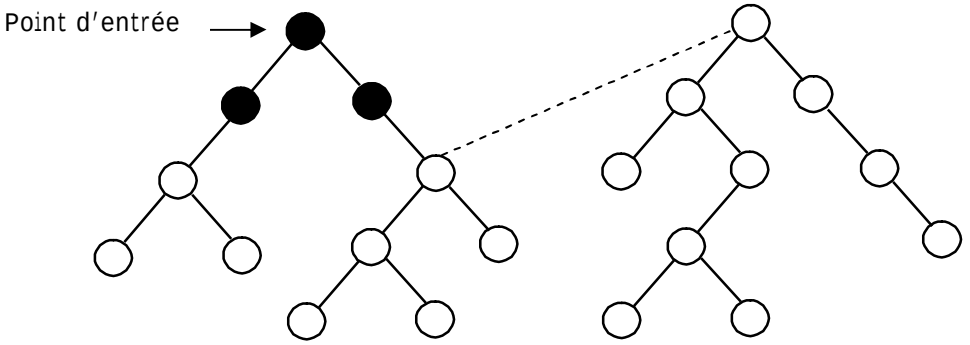
Profondeur de site : 1
Profondeur de page depuis les liens externes : 0



Cas 4 bis :
Profondeur de page : 2
Profondeur de site : 1
Profondeur de page depuis les liens externes : 1



Cas 5 :
Profondeur de page: 1
Profondeur de site : 1
Profondeur de page depuis les liens externes : 2



1.7 Stockage et acquisition de l'information

1.7.1 Stockage du contenu textuel

La troisième phase du processus de crawl consiste à stocker les informations acquises lors des deux phases précédentes.
Comme on l'a vu, le premier intérêt du parsing est d'extraire les mots porteurs de sens du texte de la page.
Par ailleurs, le code HTML d'une page est en général à 95 % dédié à la présentation et au fonctionnement de la page.
A chaque crawl d'une page Web, KB Crawl stocke systématiquement l'URL et son contenu textuel dans la base de données.
Ainsi, le moteur de recherche (cf. § 10) peut par la suite accéder directement à l'information pertinente, ce qui rend le système très performant.

NB : Tous les mots d'une même page sont stockés ensemble et dans l'ordre dans lequel ils ont été trouvés durant le parsing* avec la ponctuation qui les accompagne. Ce bloc de texte est visible pour chaque page dans l'explorateur de sources (cf. § 4).

1.7.2 Fonctions d'archivage

En plus de l'URL et de son contenu textuel, KB Crawl dispose d'un espace de stockage réservé à l'archivage au sein de sa base de données.
Lorsque KB Crawl visite une page pour la première fois, il stocke intégralement le code source de celle-ci dans cet espace.
Si des mots-clés d'alerte sont présents dans cette page, le surlignement de ces mots-clés est effectué dans la page juste avant son stockage.
Ainsi la version de référence de la page est créée.
A chaque nouvelle visite d'une page, KB Crawl compare celle-ci à sa version de référence.

Peuvent alors se présenter deux cas de figure :

1. Il n'y a aucun changement par rapport à la version de référence => La page n'est pas archivée.

2. Des changements sont détectés entre la version de référence et la version observée :

   - S'il s'agit d'une première comparaison, la nouvelle version de la page est archivée en tant que version 2.

   - Si une comparaison a déjà eu lieu, on distingue alors deux cas :

     - Cas 1 : Si le mode « mise à jour automatique » est activé dans les options de la source, alors la version 2 devient la version de référence et la page analysée est stockée en tant que version 2.

- Cas 2 : Si le mode « mise à jour automatique » n'est pas activé dans les options de la source, alors la page analysée prend la place de la version 2 et la version de référence reste la même.

Le mode « Mise à jour automatique » est indispensable lorsque l'on surveille une page dont le contenu change fréquemment afin de ne pas être alerté de manière intempestive.
En revanche, ce mode d'utilisation fait que la version de référence est écrasée et plusieurs comparaisons successives ne permettent plus de savoir quels sont les changements par rapport à la première version que l'on avait observée, sauf si l'on utilise l'option «archivage des différentes versions».

Lorsque l'option «archivage des différentes versions» est choisie, la version de référence (dans le cas 1) ou la version 2 (dans le cas 2) ne seront pas écrasées mais stockées en tant que versions intermédiaires.

Le module de gestion des archives (5) permet ensuite de suivre l'évolution dans le temps de l'information contenue pour une même URL.

2 Généralités sur l'interface

La fenêtre générale de KB crawl est composée de trois cadres distincts qui contiennent :

- la barre d'outils générale et la barre de menu textuel (cadre du haut),
- la liste de sources classées par dossier, appelée plan de classement (cadre de gauche),
- l'explorateur de sources (cadre de droite).

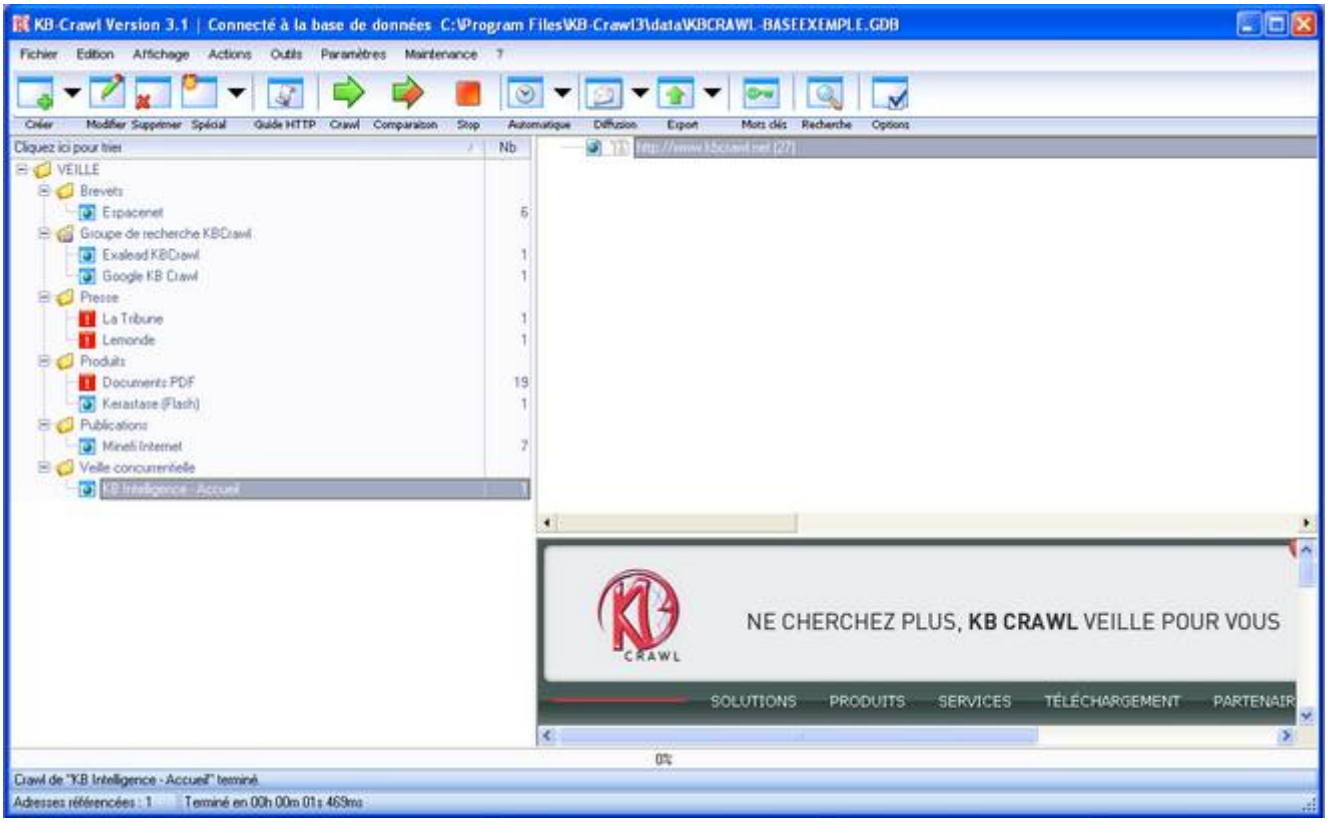


Figure 2 : Fenêtre générale de KB Crawl.

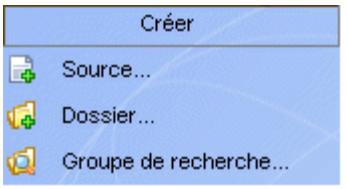
2.1 La barre d'outils générale

Figure 3 : Barre d'outils générale

La barre d'outils générale permet d'accéder directement aux fonctionnalités principales en cliquant sur ses boutons. Ces boutons, lorsqu'ils présentent une petite flèche qui pointe vers le bas à leur droite, affichent un sous menu lorsque l'on clique dessus, chaque sous menu présentant des boutons cliquables.
On peut également cliquer directement sur ces boutons pour accéder à la fonctionnalité correspondante.

De gauche à droite :

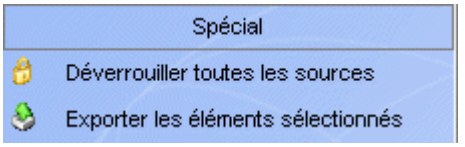
- Créer : Sous menu proposant de créer une source, un dossier ou un groupe de recherche.



- Modifier : Sous menu proposant de modifier une source ou le nom d'un dossier.

- Supprimer : Permet de supprimer une source ou un dossier ou un ensemble d'éléments sélectionnés.

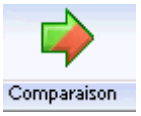
- Spécial : Sous menu qui donne accès à des fonctionnalités spécifiques : Déverrouiller toutes les sources ou exporter les éléments sélectionnés.



- Guide HTTP : Permet de gérer la liste des guides HTTP.

- Crawl : Lance un crawl d'initialisation pour les sources sélectionnées.

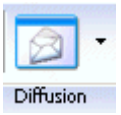
- Comparaison : Lance un crawl de comparaison pour les sources sélectionnées.



- Stop : Stoppe le crawl en cours.

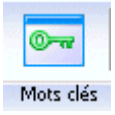
- Automatique : Sous menu qui permet d'enclencher la surveillance automatique et d'accéder au paramétrage de celle-ci.

- Diffusion : Sous menu qui permet de gérer la liste des contacts, la liste des abonnements, les e-mails en attente, et les paramètres d'envoi des e-mails d'alerte.



- Export : Sous menu qui permet de gérer les exports et les formats d'export.

- Mots-clés : Permet d'accéder au module de gestion des mots-clés.

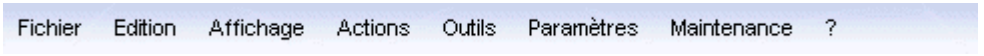


- Recherche : Permet d'accéder au module de recherche.

- Options : Permet d'accéder au menu d'options générales.

2.2 La barre de menu textuel

Elle est placée juste au-dessus de la barre d'outils générale.



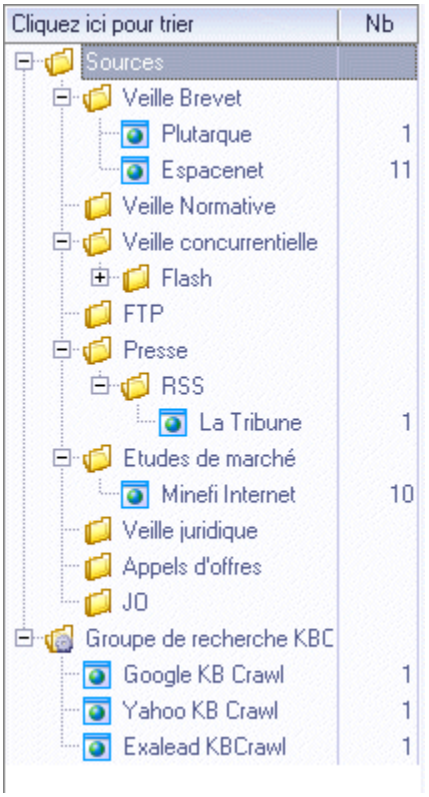
Le menu textuel permet un accès supplémentaire aux fonctions principales de KB Crawl ainsi qu'un accès à des fonctionnalités supplémentaires.
Les fonctionnalités accessibles uniquement depuis cette barre sont décrites dans le chapitre « Fonctions utilitaires» (13). La liste des fonctionnalités proposées par ce menu textuel est la suivante :

- Fichier
  - Réduire KB crawl en mode automatique
  - Quitter KB Crawl
- Edition
  - Liste des sources au format Excel
- Affichage
  - Volet de prévisualisation
  - Boite à outils URL
  - Légende
  - Journal
  - E-mail
  - KB Scraper
- Actions
  - Installer le lien KB Crawl dans Internet Explorer
  - Déverrouiller toutes les sources
  - Réinitialiser les options des sources sélectionnées
  - Supprimer les archives de la source sélectionnée
  - Initialiser toutes les connexions à la base de données
- Outils
  - Importer des favoris
  - Importer des sources venant d'une autre base
  - Import-Export des sources au format XML

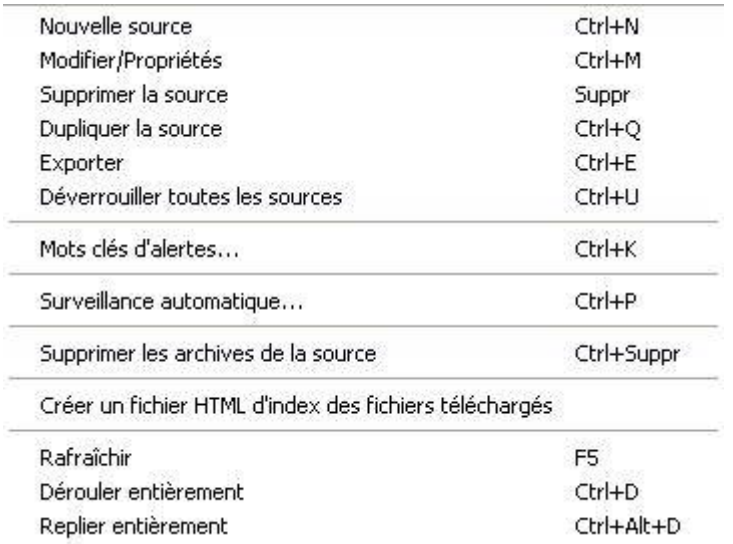
Importer des sources venant d'un fichier
Statistiques

- Paramètres
  Se connecter à une autre base de données
  Grammaire du parser
  Modifier la clé d'enregistrement KB Crawl
  Modifier la clé d'enregistrement KB Scraper
- Maintenance
  Archives
  Bases de données
  Service d'indexation
- ?
  Menu d'aide
  A propos de KB Crawl 3
  Vérifier les mises à jour
  A propos de l'éditeur

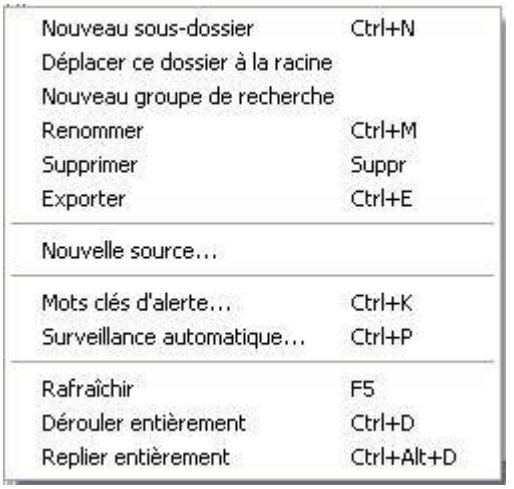
2.3 La liste des sources



Sur le côté gauche de la fenêtre principale se trouve la liste des sources présentées sous forme d'une arborescence.
Les sources sont contenues dans des dossiers ou des sous-dossiers (dossiers jaunes).
Chaque source est représentée par une icône :
Si une source fait l'objet d'une alerte (c'est à dire qu'au moins l'une des pages de cette source fait l'objet d'une alerte), l'icône qui la représente est la suivante :
La valeur dans la colonne « Nb » qui suit le libellé de la source indique le nombre de pages que celle-ci contient. Un clic droit sur une source donne accès aux fonctionnalités suivantes :



Un clic droit sur un dossier donne accès aux fonctionnalités suivantes :



NB : la plupart de ces fonctionnalités, accessibles depuis ces menus contextuels, le sont également depuis la barre d'outils générale ou par un raccourci clavier.

2.4 L'explorateur de sources

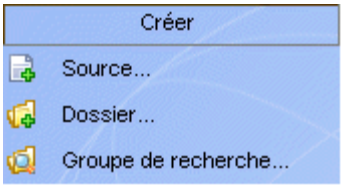
Figure 4 : L'explorateur de sources.

Ce cadre nommé « explorateur de sources » est une vue arborescente du contenu d'une source : l'arbre permet une vue hiérarchique des documents téléchargés, en fonction de l'ordre des liens, suivis lors du crawl.

3 Installation et lancement

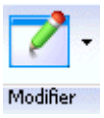
3.1 Création d'un dossier

Cliquer sur la flèche du bouton « Créer » puis dans le sous-menu, sur le bouton « Dossier … »



Saisir le nom du nouveau dossier puis valider la fiche avec la coche verte.

3.2 Modification d'un dossier



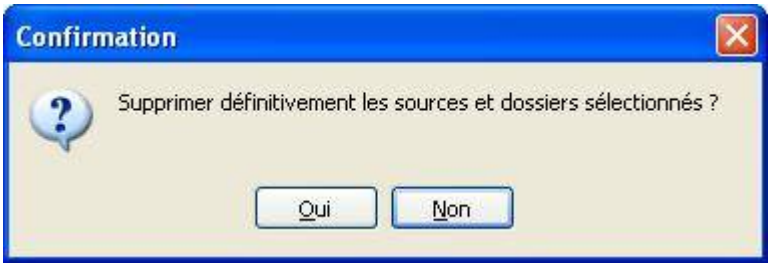
Sélectionner un dossier et cliquer sur le bouton « Modifier »

Une fois le libellé modifié, cliquer sur la coche verte pour fermer la fenêtre et enregistrer les modifications ou sur la coche rouge pour fermer la fenêtre et annuler les modifications.

3.3 Suppression d'un dossier

Sélectionner un dossier et cliquer sur le bouton « Supprimer » 

Attention : Si le dossier contient des sources ou des sous dossiers, le message suivant apparaît :



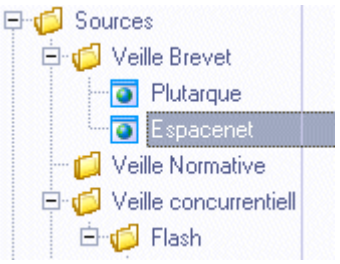
Il faut ensuite confirmer la suppression des X sources contenues dans le dossier à supprimer.

3.4 Gestion des sous dossiers

Un dossier peut contenir autant de sous dossiers voulus :

Figure 5 : Dossiers et sous dossiers

Et un dossier ou un sous dossier peut contenir autant de sources voulues :



3.4.1 Création d'un sous dossier

La création d'un sous dossier ne se fait pas depuis le même menu que les dossiers : un sous dossier doit se créer à partir d'un dossier ou d'un sous dossier afin de créer un lien de parenté à la création.
Pour créer un sous dossier, il faut sélectionner un dossier ou un sous dossier dans le cadre gauche de la fenêtre principale (liste des sources) puis faire un clic droit puis « nouveau sous-dossier »

3.4.2 Renommer un sous dossier

Se positionner sur un sous dossier et faire un clic droit puis «renommer »

3.4.3 Suppression d'un sous dossier

Se positionner sur un sous dossier et faire un clic droit puis «supprimer » (Suppr)

3.5 Ergonomie générale

Lorsque l'on crée un sous dossier ou une source, l'élément créé appartient forcément à un dossier ou un sous dossier.
Ce classement peut être modifié facilement à tout moment en utilisant les fonctionnalités classiques de « glisser-déplacer ».

On peut sélectionner une source, un groupe de sources, un ou plusieurs dossier(s)/sous dossier(s) et le(s) déplacer dans n'importe quel dossier ou sous dossier.
Dès lors qu'une source est placée dans un dossier, les propriétés (mots-clés, surveillance automatique) de ce dossier et des dossiers plus hauts dans la hiérarchie s'appliquent à cette source.

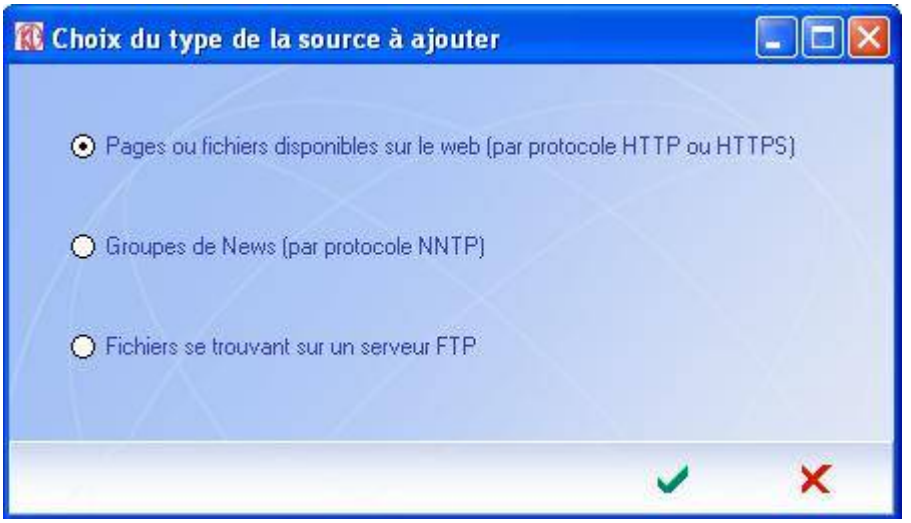
3.6 Premier crawl et paramétrage de base

Lors de la première utilisation, la page principale de KB Crawl présente une série de sources déjà paramétrées qui sont présentées à titre d'exemple.
Une source est un ensemble de pages Web dont le contenu textuel a été rassemblé puis stocké dans la base de données de KB Crawl.

Figure 6 : Liste de sources pré-paramétrées

Pour créer une source, cliquer sur le bouton « Créer » 

de la barre d'outils générale puis sur le bouton « Source », ou depuis la liste des sources, faire un clic droit puis « nouvelle source ». En cliquant sur le bouton directement, vous créez une source HTTP.



Dans KB Crawl 3, les sources se différencient selon trois types en fonction du protocole Internet utilisé pour télécharger les documents lors du crawl :

- HTTP ou HTTPS,
- NNTP,
- FTP.

Ces trois différents types de sources présentent une interface légèrement différente même si le fonctionnement général reste le même au niveau ergonomique.

Dans le bas de l'écran :

 Bouton « Valider » pour fermer la fenêtre de la source et sauvegarder tous ses paramètres.

 Bouton « Annuler » pour fermer la fenêtre de la source sans sauvegarder les paramètres éventuellement modifiés.

3.6.1 Page principale de détail d'une source

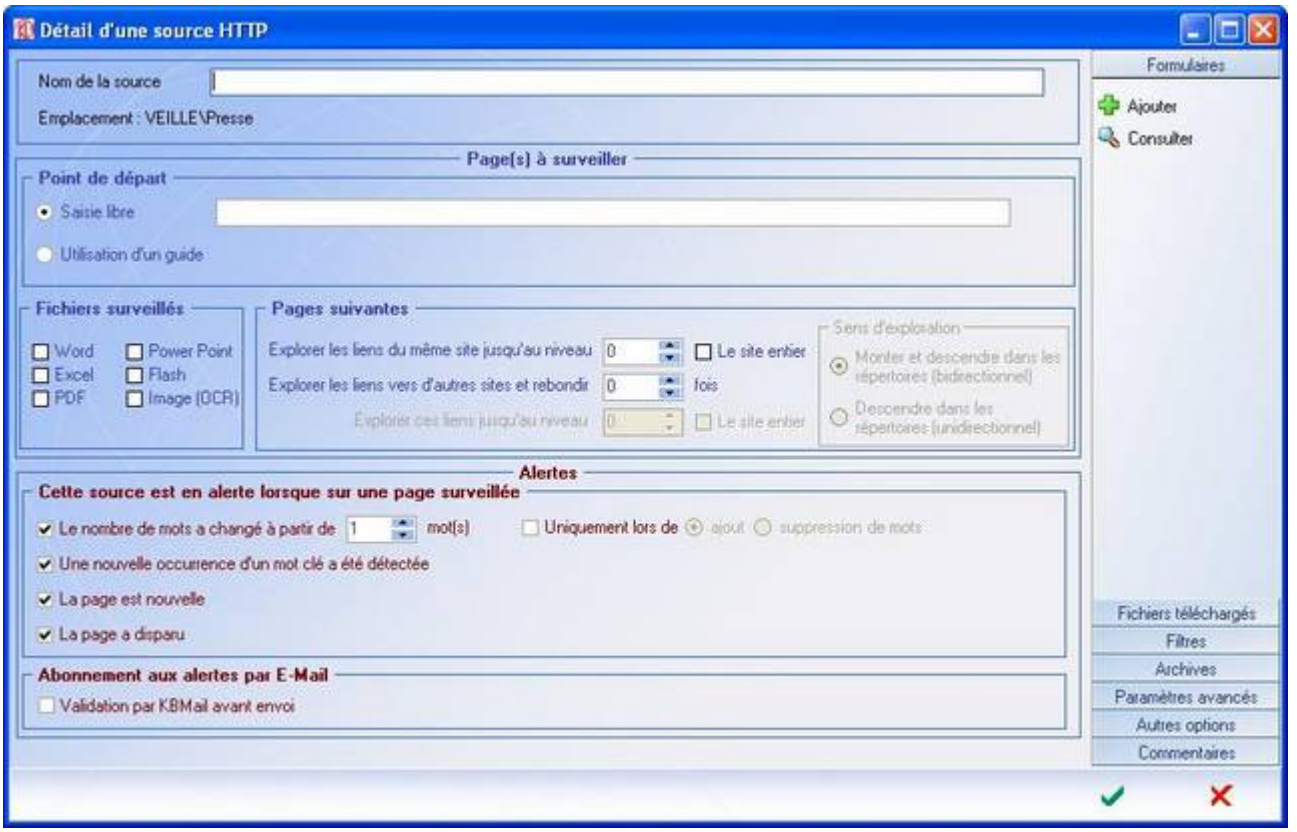


Figure 7 : Page principale d'une source.

La page principale d'une source est composée de deux cadres :

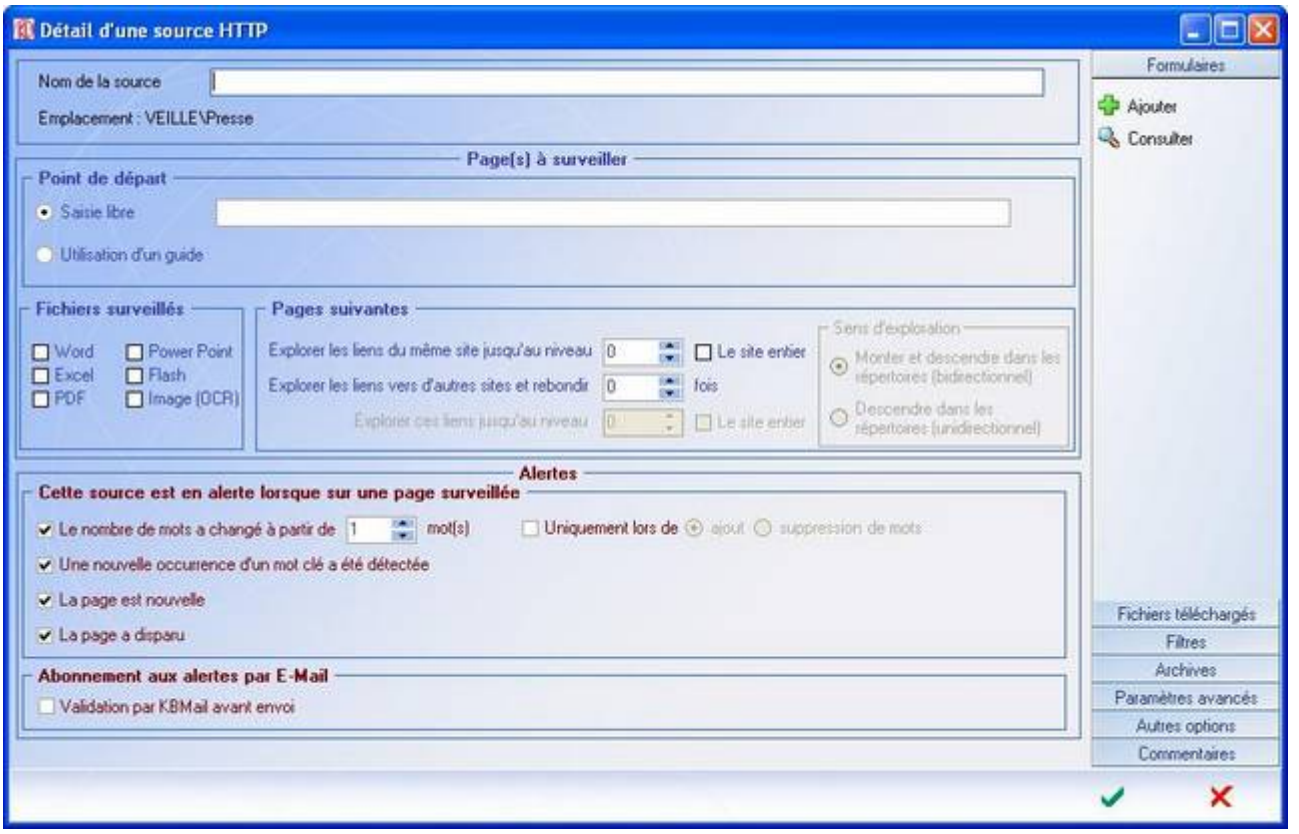
> Le cadre de gauche présente les renseignements principaux nécessaires au bon fonctionnement du crawl, tels que l'adresse de départ pour une source de type HTTP ou HTTPS, l'adresse du serveur FTP pour une source FTP, ou l'adresse du serveur NNTP pour une source NNTP.

> Le cadre de droite présente un menu à volets, qui donne accès aux paramétrages avancés de la source tels que :

- les formulaires,
- les fichiers téléchargés,
- les filtres,
- le paramétrage des archives,
- les autres options,
- les commentaires.

Le nombre de rubriques accessibles via le menu à volets dépend du type de source. Les formulaires, par exemple, sont inutiles dans les sources NNTP.

3.6.2 Source HTTP ou HTTPS



Voici de haut en bas la liste des champs qui peuvent (ou doivent) être remplis :

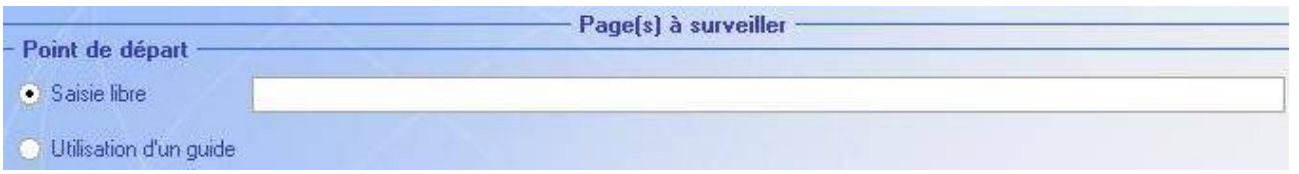
- Nom de la source

Saisir ici le nom de la source. C'est le libellé qui apparaîtra ensuite dans la liste des sources et qui permettra de la reconnaître parmi les autres.

- Point de départ

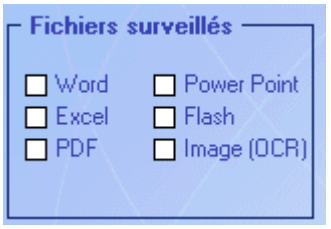
Ici, deux possibilités :

- L'URL de départ de la source est connue : Saisir ici l'adresse ou URL* complète qui définit le point de départ du crawl.
  D'une manière générale, il convient d'adopter la technique suivante : saisir cette adresse de départ dans un navigateur classique pour vérifier que l'URL est valide et qu'elle correspond réellement au point de départ souhaité. En effet, il se peut que par un jeu de redirection, l'URL correspondant au point de départ soit différente de celle considérée au premier abord. C'est alors la dernière URL indiquée par le navigateur qu'il faut saisir comme « adresse de départ» dans la source.



- Le crawl va être amené à son point de départ par un guide HTTP : on peut alors importer un fichier à l'extension « .gui » (le fichier qui matérialise le guide http). Il est également possible de créer un guide HTTP en cliquant sur le bouton « Liste des guides ».

- Fichiers surveillés



KB Crawl 3 permet de surveiller plusieurs types de formats de documents fréquemment rencontrés sur le Web : Word, Excel, PDF, Power Point, Flash et permet également d'extraire du texte contenu dans des images grâce à sa fonction OCR (voir la documentation complète du module OCR).

Afin que ces formats soient pris en compte en amont du parsing des documents, il faut cocher les cases correspondantes aux formats de fichiers.

- Surveiller les pages suivantes

Pour surveiller un site en profondeur, il est nécessaire de renseigner un certain nombre de paramètres dans le cadre « Pages suivantes » :

- Explorer les liens du même site jusqu'au niveau…

Définit la profondeur de page (§ 1.6).
Si on coche la case « le site entier », le site est alors exploré en profondeur infinie.

- Explorer les liens vers d'autres sites et rebondir « n » fois

Définit la profondeur de site (nombre de changements de nom de domaine successifs).

- Explorer ces liens jusqu'au niveau…

Définit la profondeur de page pour les sites externes visités. On peut aussi définir une profondeur de page infinie pour ces pages.

- Sens d'exploration

Sélectionner « Bidirectionnel » pour que KB Crawl explore les pages de niveaux supérieurs et inférieurs à celui de la page de départ pour l'exploration et « Unidirectionnel » pour que seules les pages de niveaux inférieurs et du même niveau que celui de la page de départ soient explorées.

Exemple :
Soit la page de départ : http://www.kbcrawl.com/products/KBCRAWL.htm
La page http://www.kbcrawl.com/actualite.html est une page de niveau supérieur à la page de départ.
En mode bidirectionnel, elle sera explorée alors qu'elle ne le sera pas en mode unidirectionnel.
Les pages http://www.kbcrawl.com/products/NiveauInferieur/page1.htm et http://www.kbcrawl.com/products/KBCrawl_2.htm seront toutes deux explorées dans les deux cas.

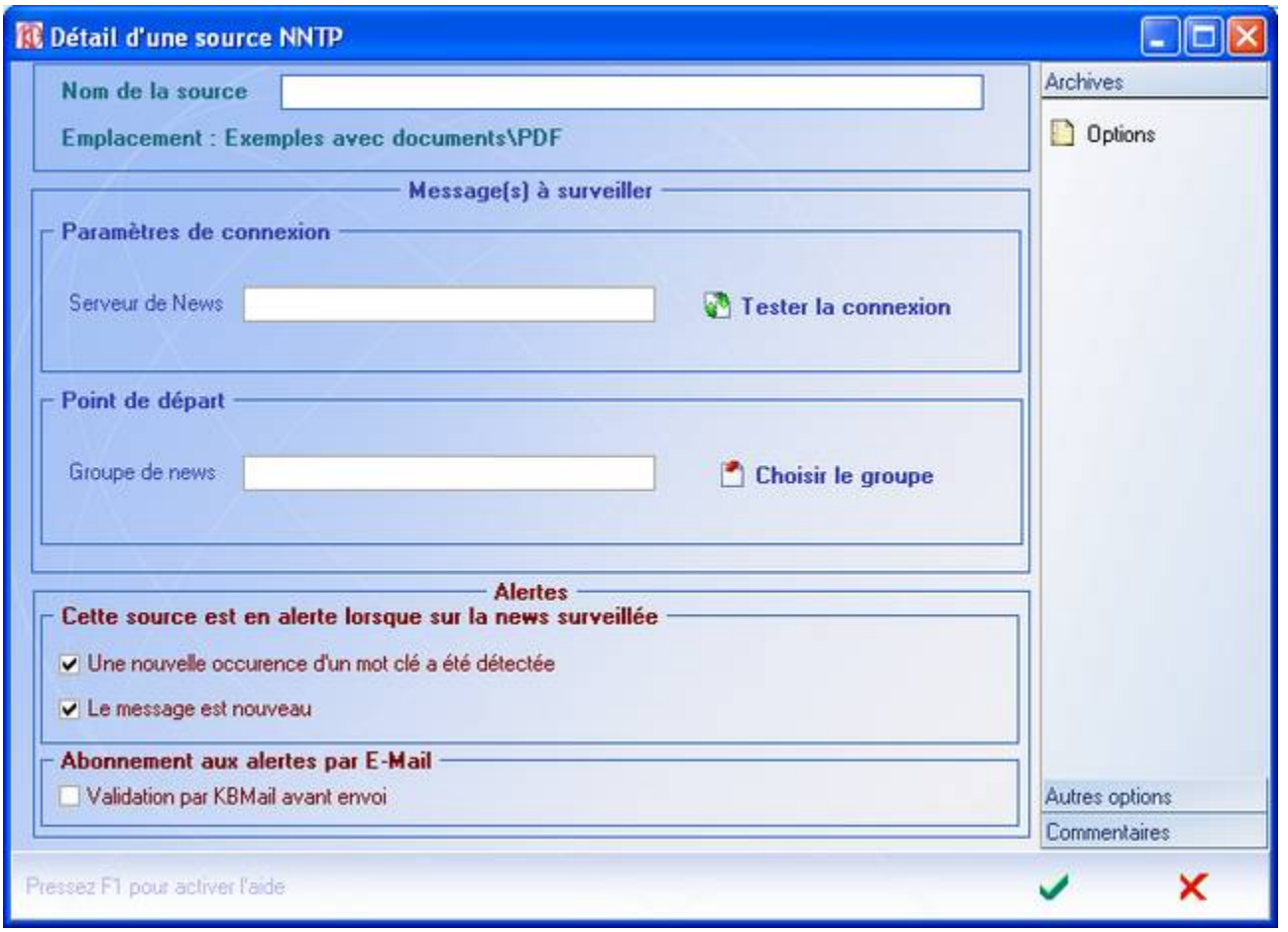
- Alertes

Cocher ici les critères qui doivent déclencher une alerte une fois qu'un crawl de comparaison a été réalisé.

- Abonnement aux alertes par E-Mail

Cette section ne concerne que la validation des e-mails d'alerte avant l'envoi et n'est utile que si ceux-ci ne sont pas envoyés automatiquement après chaque crawl.
Les e-mails qui requièrent une validation seront à valider avant envoi dans le module KB Mail si celui-ci est utilisé.

3.6.3 Source NNTP



- Nom de la source

Identique à la source HTTP ou HTTPS

- Paramètres de connexion

Un champ est prévu à cet emplacement pour saisir le nom du serveur de News.
Le bouton « Tester la connexion » sert ensuite à vérifier la connexion avec le serveur de news.

- Point de départ

Saisir le nom du groupe de news. Celui-ci doit être référencé par le serveur de groupe de news. Pour obtenir la liste des groupes disponibles sur ce serveur, cliquer sur le bouton « Choisir le groupe ».

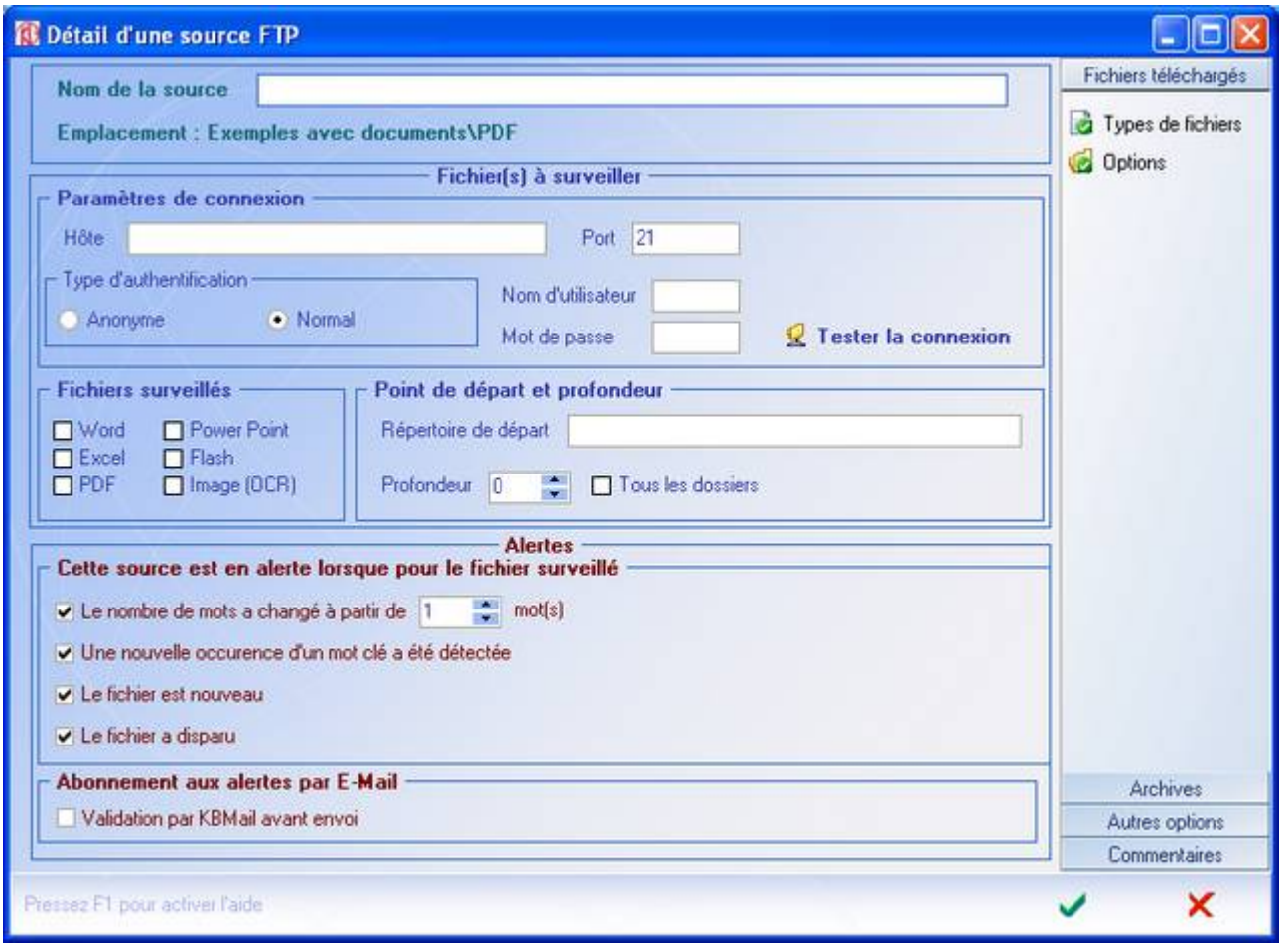
- Cette source est en alerte lorsque sur la news surveillée

Choisir ici les critères qui déclenchent l'alerte.

- Abonnement aux alertes par E-Mail

Identique à la source HTTP ou HTTPS

3.6.4 Source FTP



- Nom de la source

Identique à la source HTTP ou HTTPS

- Paramètres de connexion

- Hôte : saisir ici l'adresse du serveur FTP
- Port : saisir ici le port du serveur FTP (généralement 21)
- Type d'authentification : choisir entre anonyme ou normal
- Nom d'utilisateur : saisir ici le nom d'utilisateur pour se connecter au serveur FTP
- Mot de passe : saisir ici le mot de passe pour se connecter au serveur FTP

Une fois ces paramètres saisis, il est possible de tester la connexion grâce au bouton « Tester la connexion ».

- Fichiers surveillés

Identique à la source HTTP ou HTTPS

- Point de départ et profondeur

- Répertoire de départ : saisir ici le chemin de départ de l'exploration sur le serveur FTP
- Profondeur : saisir ici la profondeur d'exploration dans les répertoires du serveur FTP ou bien cocher la case « Tous les dossiers » pour explorer tous les répertoires descendants de celui de départ. Pour définir la racine du serveur comme départ, saisir « / ».

- Cette source est en alerte lorsque pour le fichier surveillé

Identique à la source HTTP ou HTTPS

- Abonnement aux alertes par E-Mail

Identique à la source HTTP ou HTTPS

Ces paramétrages de base suffisent à lancer un premier crawl, sans qu'il soit nécessaire d'aller dans les paramètres avancés. Il est même recommandé, pour la création d'une source dans KB Crawl, de fonctionner par étapes et de s'en tenir à ces paramétrages de base pour un premier crawl.

Il suffit alors de valider la fiche de la source en cliquant sur l'icône « Valider » : 

KB Crawl propose alors de lancer un crawl d'initialisation :

3.7 Paramétrage avancé d'une source

3.7.1 Menu Formulaires

3.7.1.1 Introduction

On appelle «formulaire» l'ensemble du code HTML situé dans le code source d'une page et placé entre deux balises : <FORM> où débute le code du formulaire et </FORM> où il se termine.
Le navigateur Web interprète ce code et produit une interface pour l'utilisateur afin que celui-ci puisse saisir un certain nombre de données. La saisie de ces données se fait grâce à des zones de saisie libre, des listes déroulantes, des boutons radios ou des cases à cocher.
Un formulaire est généralement accompagné d'un bouton cliquable dont le libellé varie.
On retrouve cependant fréquemment les libellés suivant « Envoyer » ou « Rechercher ».
Lorsqu'on appuie sur ce bouton, les données renseignées par l'utilisateur sont envoyées à un serveur Web dont l'adresse est inscrite dans le code du formulaire. Le serveur Web répond ensuite à l'internaute en fonction des données qu'il a reçues.

On retrouve très fréquemment deux types de formulaires dont voici deux exemples :

- le formulaire d'authentification :

Figure 8 : Exemple de formulaire Web d'authentification.

Différente de l'authentification de base, l'authentification par formulaire est intégrée dans la page Web. Sa forme varie à l'infini selon l'environnement graphique de chaque site Internet.
Suite à l'envoi des données par ce formulaire, on obtient généralement une page qui montre que l'on s'est authentifié correctement et que l'on a accès au site Internet sécurisé ou bien une page qui exprime le refus d'accéder au reste des pages et éventuellement qui invite à retenter l'authentification.

- le formulaire de moteur de recherche

Figure 9 : Exemple de formulaire Web de moteur de recherche.

Suite à l'envoi des données de ce formulaire, le serveur Web qui reçoit la requête construit une page de résultats et l'envoie comme réponse au navigateur Web.
C'est ainsi que fonctionnent tous les moteurs de recherche sur le Web.

Ainsi, une grande partie des informations disponibles sur le Web est « cachée » derrière ces formulaires et les systèmes de sécurité mis en place empêchent de télécharger une page directement sans les avoir correctement remplis.
KB Crawl permet d'automatiser cette tâche afin de récupérer et surveiller les informations de ces sites sécurisés.
Pour se faire, il est nécessaire que KB Crawl « sache » quelles données envoyer à un formulaire lorsqu'il le rencontre au cours d'un crawl, d'où la nécessité d'enregistrer au préalable ces données rattachées à un formulaire.
Pour une même source, on peut enregistrer autant de formulaires que l'on souhaite grâce à l'analyseur de formulaires.

Le menu « Formulaires » propose deux fonctionnalités :

3.7.1.2 *Ajouter un ou plusieurs formulaires*

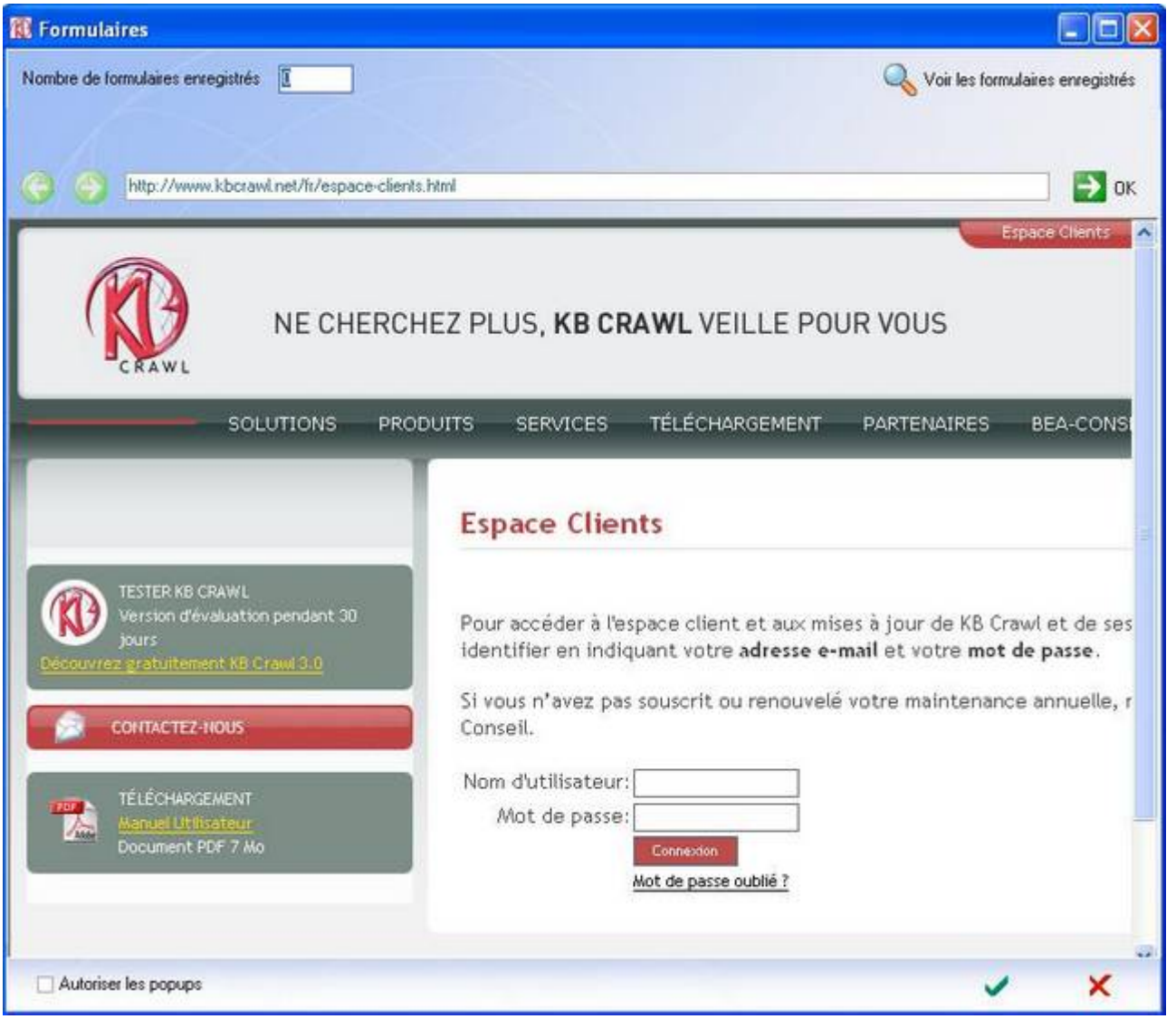


Figure 10 : L'analyseur de formulaires.

L'analyseur de formulaires est un navigateur Web intégré à l'interface de KB Crawl.
Il présente une zone de saisie libre dans laquelle on peut saisir une URL.
Pour naviguer sur la page requise, il suffit d'appuyer sur la touche <ENTREE> du clavier ou de cliquer sur bouton « OK » .
Dans l'exemple ci-dessus, on navigue sur un site qui exige une authentification.
L'analyseur de formulaires permet de « surfer » sur Internet exactement comme le navigateur utilisé par défaut. On peut cliquer sur des liens hypertexte, mais aussi, et surtout, remplir des formulaires.

Lorsque l'on clique sur le bouton « Envoyer » de la page Web vue à l'intérieur de l'analyseur de formulaires, celui-ci détecte cette action automatiquement et propose alors d'enregistrer le formulaire et de l'ajouter à la liste des formulaires enregistrés pour cette source :

Figure 11 : L'analyseur de formulaire détecte l'envoi de données.

Si l'on confirme avec le bouton « Oui », les données du formulaire seront enregistrées dans la base de données de KB Crawl et aucun paramétrage supplémentaire n'est nécessaire.

Autre exemple :

Figure 12 : L'analyseur de formulaires avec un moteur de recherche.

(Source : Google)

On navigue sur la page d'accueil d'un moteur de recherche, puis on saisit une requête dans la zone prévue à cet effet.
Ensuite, on clique sur le bouton qui sert à déclencher la recherche et on enregistre le formulaire.
Cette opération peut être ainsi répétée autant de fois qu'on le souhaite pour un même formulaire :

Dans le cas d'enregistrement de données multiples programmées pour un même formulaire, KB Crawl enverra autant de requêtes différentes pour le même formulaire et enregistrera chaque page de résultats du serveur distant comme une page à part entière du site (voir plus bas : exemples de crawl avec formulaires).

3.7.1.3 *Voir les formulaires enregistrés*

Chaque formulaire enregistré est ajouté à une liste rattachée à la source dont le nombre d'éléments n'est pas limité.
L'analyseur de formulaires de KB Crawl est conçu pour que la technique sous-jacente aux formulaires demeure transparente et que leur manipulation reste simple afin de s'affranchir complètement de la partie technique liée aux formulaires.
Toutefois, on peut, si on le souhaite, gérer soi-même la liste des formulaires depuis l'interface de KB Crawl, ces manipulations relevant d'une utilisation avancée.
Dans le cadre de gauche, une grille montre la liste des formulaires enregistrés (voir plus haut Figure 10 : L'analyseur de formulaires). Lorsque l'on pointe sur une ligne de la grille, le cadre de droite fait apparaître le formulaire tel qu'il a été enregistré au format HTML.

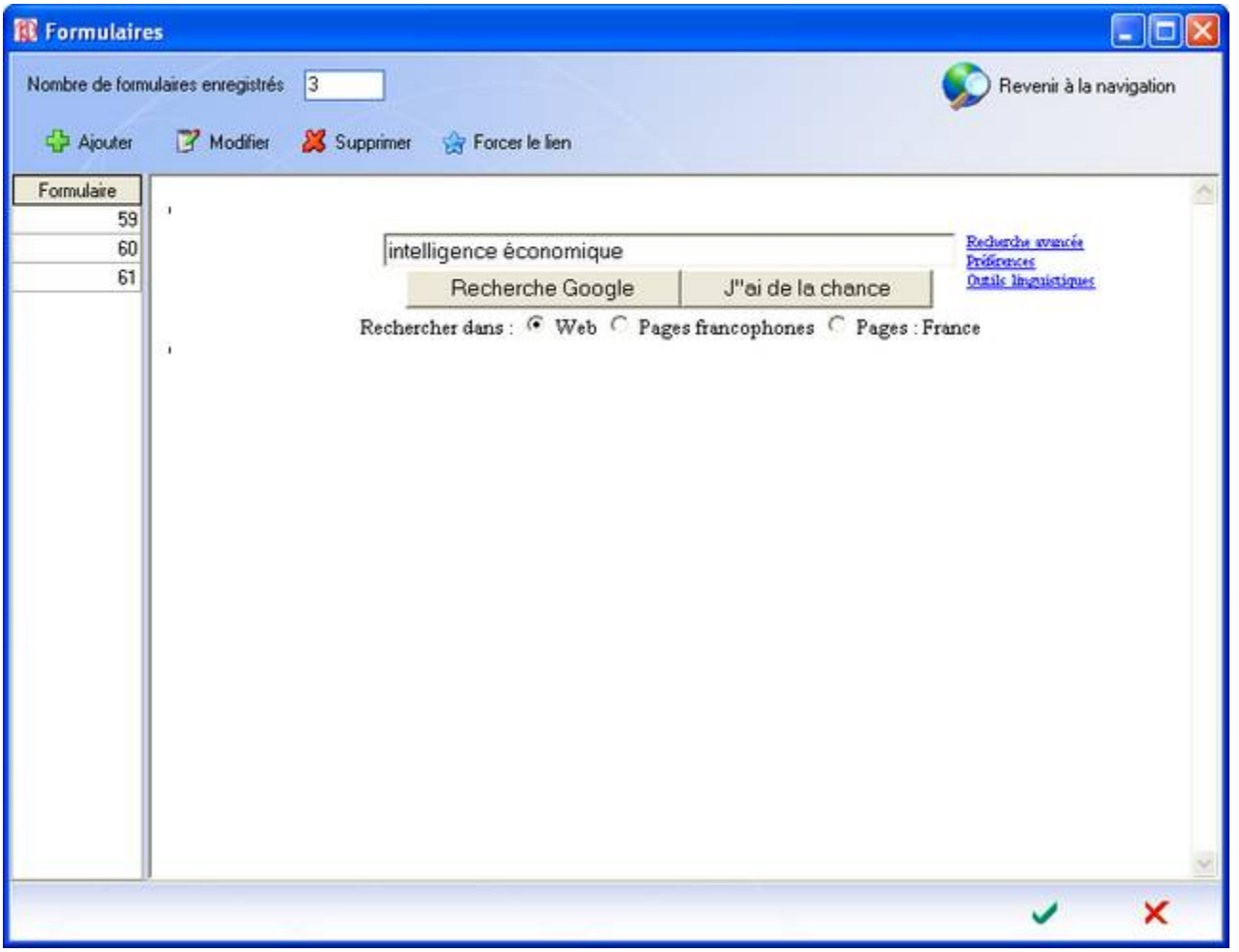
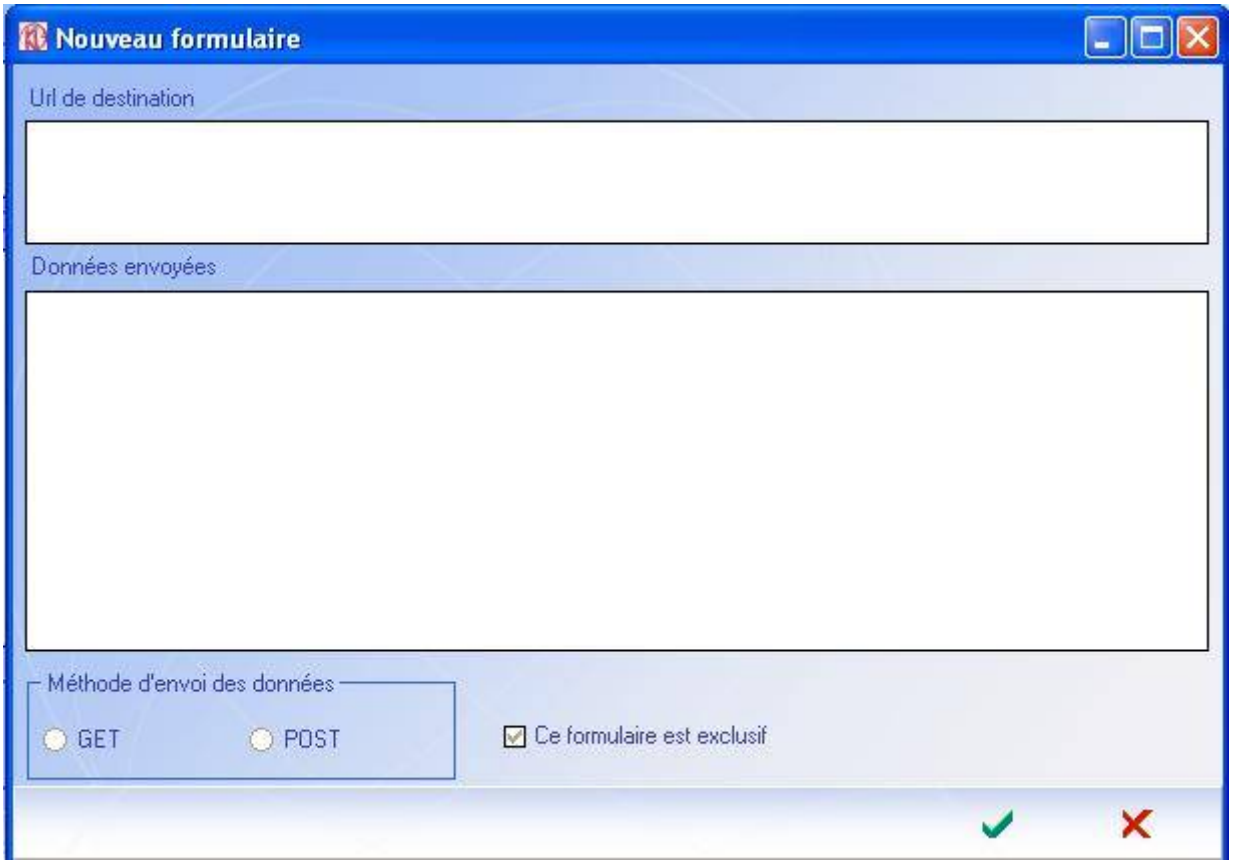


Figure 13 : Liste des formulaires enregistrés.

- Ajouter un formulaire

Cliquer sur le bouton « Ajouter » puis renseigner les champs présents dans cet écran :



- Url de destination : libellé de l'URL qui correspond à la variable « ACTION » que l'on trouve dans le code source du formulaire.

- Données envoyées : chaine de caractères encodée au format HTTP qui réunit l'ensemble des données envoyées au formulaire (sous la forme ?PARAM1=VALEUR1&PARAM2=VALEUR&PARAM3=VALEUR3 etc.)

- Méthode d'envoi des données : cliquer sur le bouton radio « GET » ou « POST » selon la méthode d'envoi des données au serveur. La méthode d'un formulaire se trouve en principe dans son code source et correspond à la variable « METHOD »

- Option : « Ce formulaire est exclusif» : cette option est très importante ; lorsque l'on se situe sur une page web contenant un formulaire, cette page contient le lien pour l'adresse de destination du formulaire (« URL de destination ») mais aussi bien souvent d'autres liens. On ne souhaite en général pas que KB Crawl suive ces autres liens mais plutôt qu'il se concentre sur le formulaire en question.

- Modifier un formulaire

Cliquer sur le bouton « Modifier », puis modifier les champs proposés à cet effet

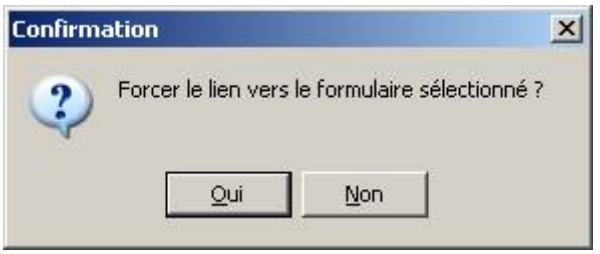
- Supprimer un formulaire

Cliquer sur le bouton « Supprimer »

- Forcer le lien

Cliquer sur le bouton « Forcer le lien» 

La boite dialogue de confirmation apparaît :



Un clic sur « Oui » fait apparaître le message suivant :

Certains formulaires ne font pas figurer leur adresse de destination dans la balise <ACTION> qui leur correspond (§ 3.7.1.1).
Ainsi, il se peut que le module de parsing de KB Crawl (§ 1.5) ne puisse pas retrouver lors d'un crawl un formulaire qui a été enregistré.
Si ce cas se produit, il suffit d'ajouter l'adresse du formulaire « caché » à la « liste des liens forcés ». Lors du processus de crawl, à chaque niveau d'arborescence, KB Crawl visite les pages correspondantes aux liens forcés.
Pour cette raison, il est important de préciser à quel niveau de l'arborescence le lien est à visiter. Par défaut, il est créé pour le niveau 0 de l'arborescence comme l'indique la boite de dialogue précédente.

NB : Toutes ces actions ne sont pas enregistrées tant que l'on n'a pas cliqué sur le bouton « Valider » de la fenêtre de propriété de la source.

3.7.1.4 Le crawl avec des formulaires

Lorsque KB Crawl analyse une page pour en extraire les liens, il analyse le code des éventuels formulaires (option cochée par défaut dans l'onglet « Options »). Lors de l'analyse d'un formulaire, il en extrait l'adresse de destination et l'ajoute à la liste des URL trouvées pour cette page.
Ensuite, si la profondeur de page le permet (cf. § 1.6), il va comparer cette adresse de destination à celles qui sont enregistrées dans la liste des formulaires. Si l'une des adresses de ces formulaires correspond à cette adresse de destination, KB Crawl envoie à cette adresse les données enregistrées pour ce formulaire avec la méthode adéquate (GET ou POST).

La page renvoyée ensuite par le serveur après qu'il ait reçu ces données est marquée dans la base de données de KB Crawl comme une page de type formulaire.
Ainsi, on peut retrouver facilement cette page dans l'explorateur de sources car cette dernière est représentée par une icône particulière :

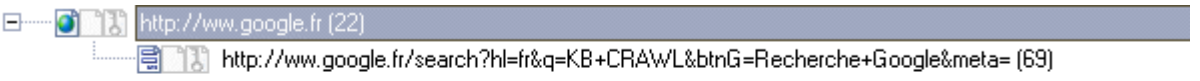


Figure 14 : Repérage d'un formulaire dans l'explorateur de sources

Sur cet exemple, on voit le résultat du crawl du formulaire d'identification.
La première URL (http://www.google.fr) contient le formulaire.
La seconde URL (http://www.google.fr/search?hl=fr&q=KB+CRAWL&meta=) est l'adresse de destination du formulaire contenu dans la première page qui est la page mère.
Pour obtenir ce résultat, il est donc nécessaire de paramétrer pour cette source une profondeur de page de 1 :

Autre exemple :

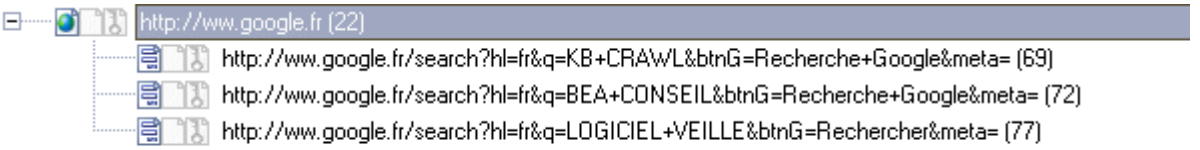


Figure 15 : Formulaires multiples dans l'explorateur de sources

Ici, l'URL de départ (www.google.fr) contient un formulaire servant aux requêtes de recherche.
Pour la source correspondante, on a enregistré trois jeux de données différents pour le même formulaire afin d'effectuer trois recherches différentes.
Ainsi, dans l'explorateur de sources, on voit que les résultats de chaque requête sont matérialisés par une page différente symbolisée par l'icône « formulaire ».
Comme dans l'exemple précédent, la source est paramétrée avec une profondeur de page de 1.
Si on paramètre une profondeur de page supérieure, chaque page résultant de la requête de recherche engendrera un certain nombre de pages filles.

3.7.2 « Fichiers téléchargés »

3.7.2.1 *Fichiers à enregistrer*

Lors d'un crawl, KB Crawl détecte les liens qui mènent à des fichiers (images, fichiers Excel, PDF ou médias divers, MP3, vidéos, etc.).
Par défaut, KB Crawl ignore ces fichiers et ne télécharge que ceux dont le format est texte/HTML.
Ceci permet d'optimiser le temps d'exploration en évitant de télécharger des fichiers volumineux qui ne contiennent pas de texte.
Cependant, KB Crawl peut également récupérer ces fichiers additionnels en ajoutant ces extensions à la bibliothèque d'extensions.

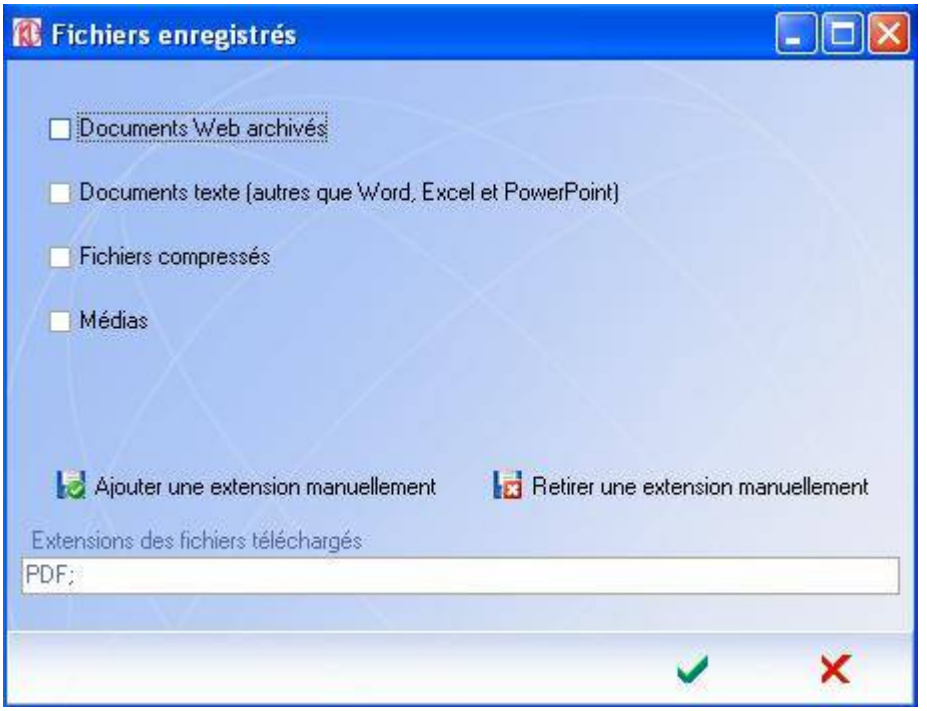
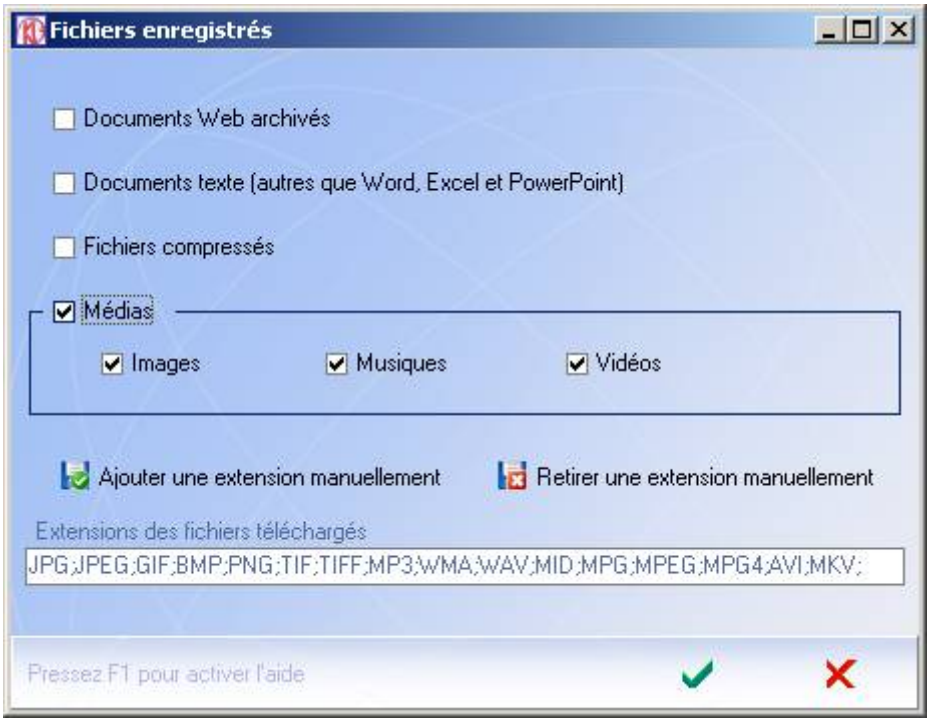


Figure 16 : Onglet "Fichiers à enregistrer" du détail d'une source.

Dans le cas ci-dessus, on souhaite récupérer les fichiers PDF.
L'extension PDF apparaît car, dans la fiche principale, la case correspondante à l'extraction des fichiers PDF est cochée.
Le menu «Fichiers enregistrés » permet d'ajouter d'autres extensions automatiquement, en cochant l'une des cases suivantes : « Documents texte (autre que Word, Excel et PowerPoint) », « Fichiers compressés » ou « Médias », qui se subdivisent en sous types :

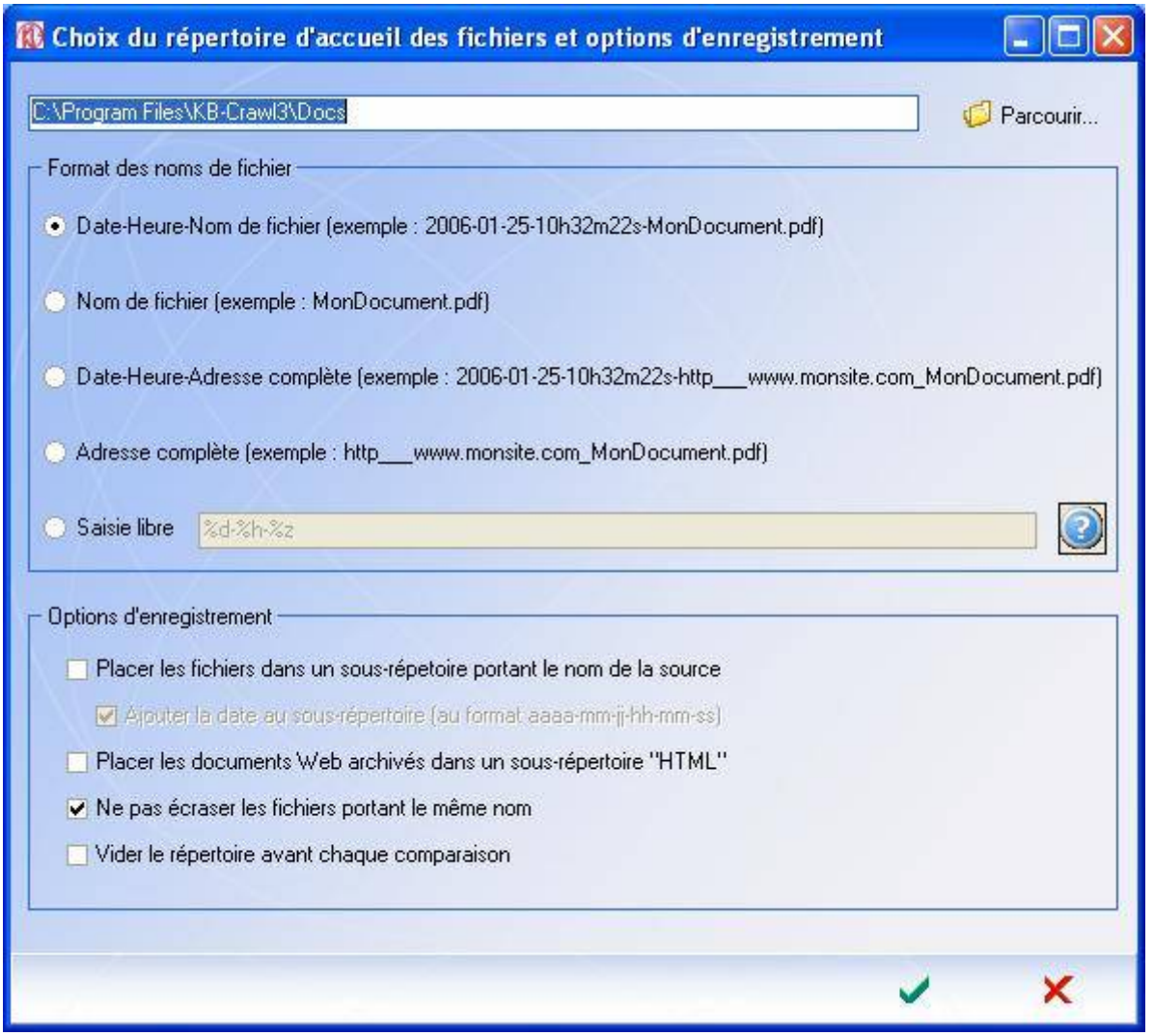


On peut également ajouter une extension manuellement :

Ou encore en supprimer une de la même manière :

3.7.2.2 *Répertoire d'accueil des fichiers*

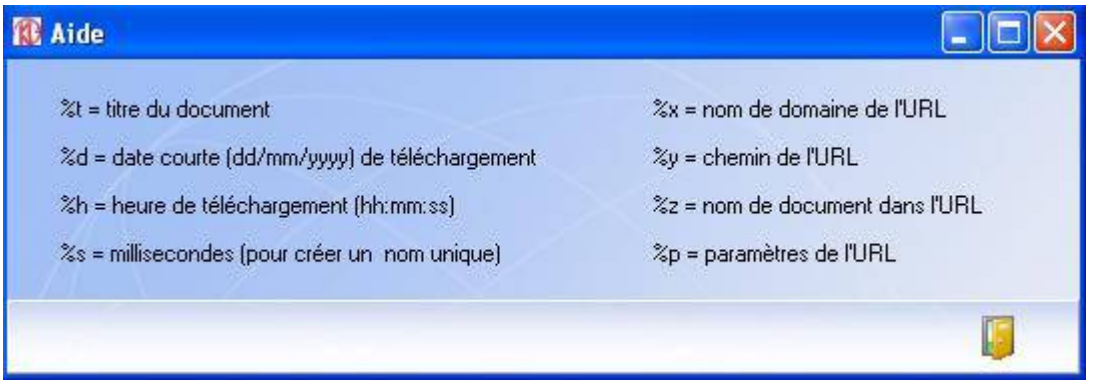
Les documents collectés sont placés dans le répertoire que l'on a désigné comme réceptacle dans le champ intitulé « répertoire d'accueil des fichiers ».



Ce répertoire est désigné dans le premier champ de la fiche ; il est possible de sélectionner un autre répertoire que celui proposé par défaut à l'aide du bouton « parcourir ».

Ensuite, il est possible de spécifier le format du nom des fichiers enregistrés. Des formats préétablis sont proposés (4 premiers boutons radios) car ils sont pratiques et souvent utilisés ; Il est également possible de composer son propre masque de nom de fichier à partir de variables comme %d par exemple qui prendra la valeur de la date au moment de l'enregistrement du fichier.

Un bouton d'aide permet de voir la liste des variables que l'on peut utiliser :

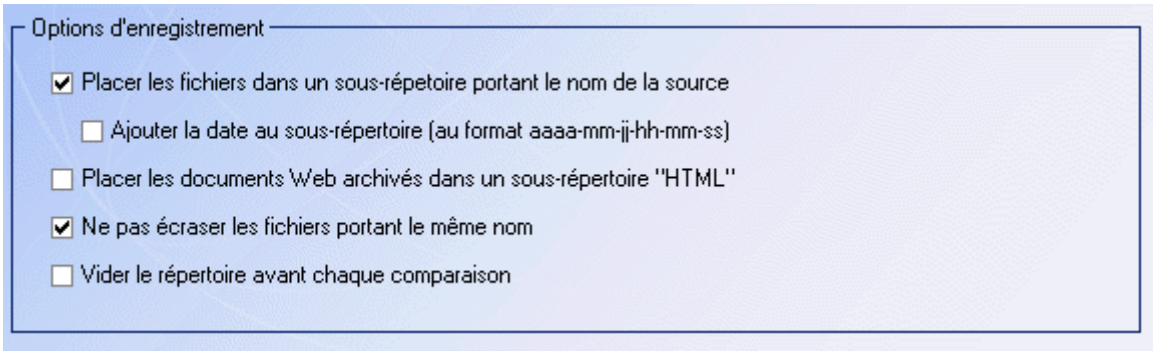


Exemple de masque : %x_%t _%d_%h pourra donner le nom de fichier suivant :

www.kbcrawl.com_presentation.pdf_15_01_2006_15_05_55

Notons que les caractères « / » et « : » ont été remplacés automatiquement par des « _ » parce qu'ils sont interdits dans les noms de fichiers Windows.

Il est ensuite possible de spécifier les options d'enregistrement.
Par défaut, chaque document est placé dans un sous répertoire portant le nom de la source à laquelle il est rattaché et les documents s'accumulent dans le répertoire au fur et à mesure des crawls :



3.7.3 Filtres

Pour accéder au paramétrage des filtres depuis la fiche de détail de la source, cliquer sur le bouton filtre du menu à volets situé sur la droite. L'écran de paramétrage des filtres s'ouvre :

Lorsque l'on paramètre une profondeur de pages supérieure à zéro dans une source, KB Crawl explore un ensemble de pages et constitue ensuite une arborescence représentant la hiérarchie des pages entre elles (4).
Sur l'ensemble des pages trouvées par KB Crawl, il se peut que seul un sous-ensemble de ces pages soit intéressant à surveiller.
Il devient alors intéressant de « filtrer » l'ensemble des pages trouvées par KB Crawl afin de déterminer un périmètre de scrutation restreint et personnalisé.
Pour cela, il est possible et souhaitable d'appliquer deux types de filtrage :

3.7.3.1 Exclusivité

Lorsqu'une URL est placée dans le filtre « Exclusivité », elle devient exclusive pour un niveau de crawl donné : lorsque cette URL est trouvée à l'intérieur d'une page, elle est explorée de façon exclusive, en évitant toutes celles qui sont à son niveau.

3.7.3.2 Black List

Pour éviter de crawler une URL, il suffit de la « black lister », ainsi, lorsque KB Crawl la rencontre, il la reconnaît en tant que telle et ne la crawle pas.

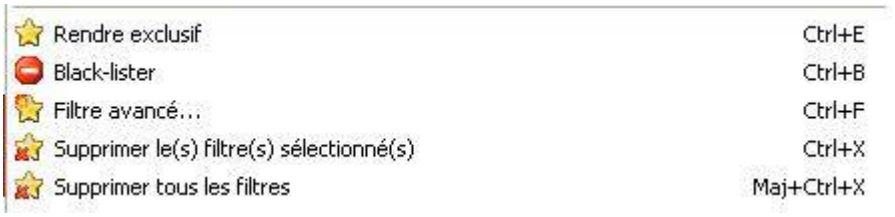
3.7.3.3 Liens forcés

Lorsque le Parser de KB Crawl ne peut trouver un lien qui doit le mener d'une page 1 vers une page 2 et que l'on souhaite nécessairement visiter la page 2 parce qu'elle contient des informations intéressantes, il suffit de créer un lien forcé à partir du niveau d'arborescence de la page 1.
La page 2 ne sera pas forcément rattachée à la page 1 si cette page n'est pas unique à son niveau d'arborescence car le lien vers la page 2 n'a pas été trouvé dans la page 1 mais forcé depuis le niveau d'arborescence de la page 1.

3.7.3.4 Ajouter un filtre

Pour rendre une URL exclusive ou la black-lister, le moyen le plus direct et le plus simple est l'explorateur de sources. Il faut d'abord sélectionner dans l'explorateur la ou les URL sur lesquelles on souhaite appliquer un filtre.

Ensuite, il est possible de faire un clic droit sur la ou les URL sélectionnées et de cliquer sur le bouton du menu contextuel correspondant au filtre souhaité.



Il est également possible d'utiliser la boite à outils en sélectionnant le filtre adapté.

Aussitôt appliqué, le filtre est visible depuis l'explorateur de sources :

Exemple 1 : URL black-listée

Figure 17 : Filtre de type "black-liste" visible depuis l'explorateur de sources.

Exemple 2 : URL exclusive

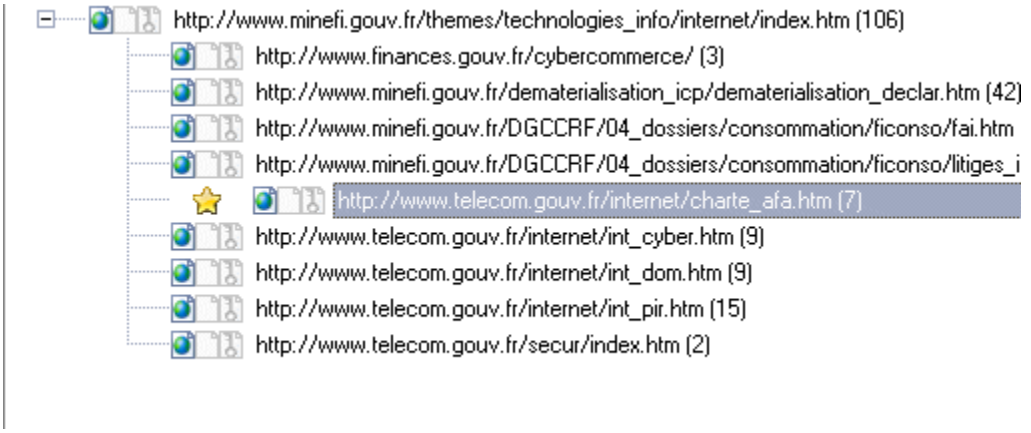


Figure 18 : Filtre de type "Exclusif" visible depuis l'explorateur de sources.

Il est possible de sélectionner plusieurs URL à la fois pour leur appliquer un même filtre :

Figure 19 : Application d'un filtre à plusieurs URL simultanément.

Parmi les autres fonctionnalités concernant le filtrage accessibles depuis l'explorateur de sources, il est possible de supprimer les filtres des URL sélectionnées (SUPPR), ou encore supprimer tous les filtres d'une source (CTRL+SUPPR).

Suite à l'application d'un ou plusieurs filtres sur une URL, ceux-ci apparaissent dès que l'on consulte le menu filtre depuis la source :

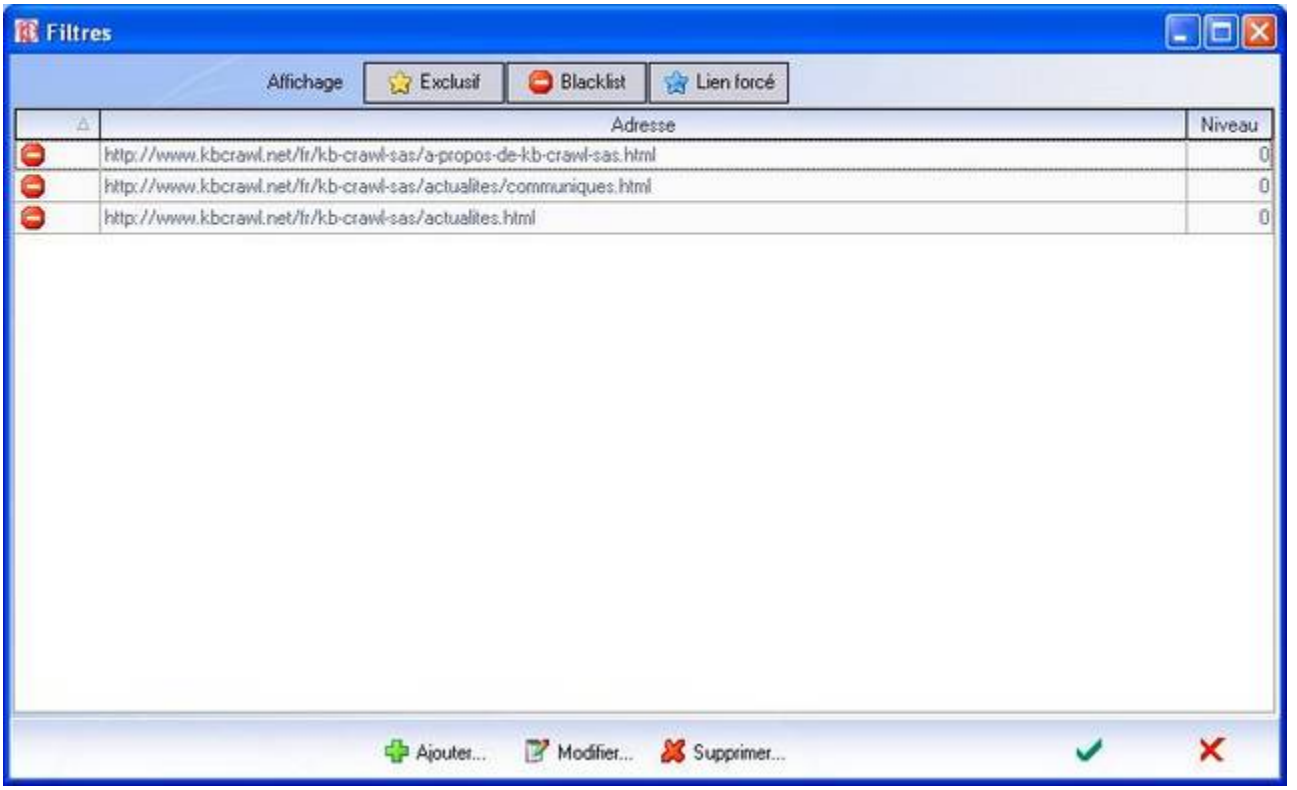


Figure 20 : Onglet "Filtre" du détail d'une source.

L'entier qui apparaît à droite de l'URL (ici égal à zéro) est le niveau d'arborescence auquel ce filtre s'applique. Lorsque cet entier est -1, le filtre s'applique à tous les niveaux d'arborescence de la source.

Depuis l'explorateur de sources, un assistant permet de créer des filtres plus complexes qui permettent au module de parsing (1.5) de répondre à des besoins plus pointus pour le filtrage des URL.
Pour accéder à cet assistant, il suffit d'un clic droit + « Filtres avancés » (ou CTRL+F)

L'assistant de filtres avancés permet de définir :

- le type de filtre : « Exclusif » ou « Blacklist »,
- Le niveau d'application du filtre : pour que le filtre s'applique à tous les niveaux, cocher la case correspondante à cette option,
- éventuellement un filtre avec un paramètre variable.

3.7.3.5 Filtre à paramètre variable

Certaines URL trouvées sur une page Web sont semblables, à un paramètre près.
Exemple :
http://www.google.fr/search?q=KB+CRAWL&hl=fr&lr=&ie=UTF-8&oe=UTF-8&start=10&sa=N
et
http://www.google.fr/search?q=KB+CRAWL&hl=fr&lr=&ie=UTF-8&oe=UTF-8&start=20&sa=N

Seule la valeur du paramètre « start » différencie ces deux URL.

Figure 21 : Liens sur les résultats de recherche Google

Ces deux URL correspondent aux liens vers les pages de résultats « 2 » et « 3 » de Google.

Soit la problématique suivante : on souhaite crawler tous les résultats d'une recherche de Google en récupérant exclusivement les pages qui correspondent à ces résultats.

Si l'on paramètre une profondeur de page supérieure à 1 pour que KB Crawl analyse les pages correspondantes aux liens vus plus haut, il va récupérer au passage beaucoup de pages parasites et les black-lister toutes s'avère fastidieux.
On ne peut pas non plus rendre exclusive l'URL suivante :
http://www.google.fr/search?q=KB+CRAWL&hl=fr&lr=&ie=UTF-8&oe=UTF-8&start=10&sa=N
car dans ce cas, les autres pages de résultats ne seront pas prises en compte.

La solution est de créer un filtre exclusif à paramètre variable :

http://www.google.fr/search?q=KB+CRAWL&hl=fr&lr=&ie=UTF-8&oe=UTF-8&start=[*]&sa=N

La valeur du paramètre qui varie doit être remplacée par [*]
Ainsi, toutes les URL dont seules la valeur du paramètre « start » est différente deviennent exclusives pour un niveau donné ou bien tous les niveaux.
Il n'est pas toujours simple à l'œil nu de repérer le paramètre variable d'une URL, c'est pour cela que l'assistant de filtres avancés permet de le générer automatiquement.

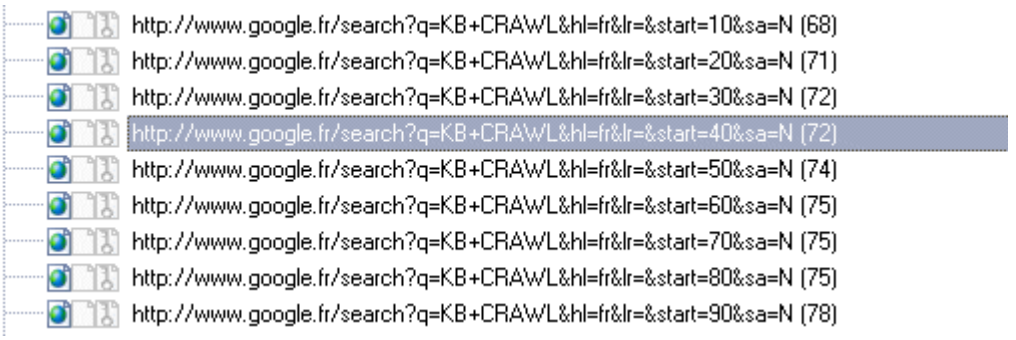
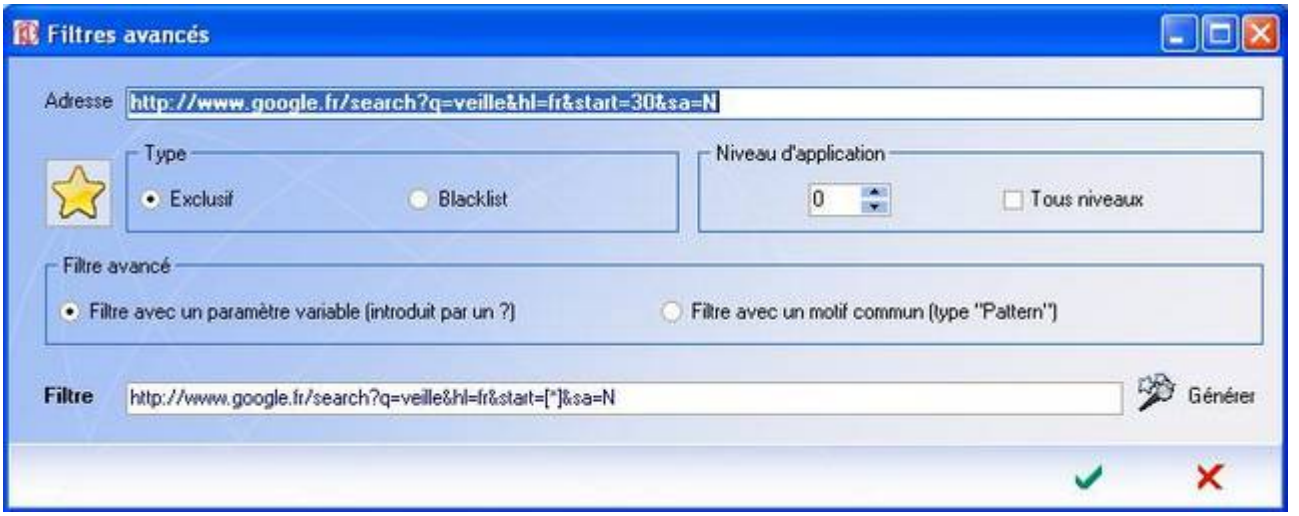


Figure 22 : URL à paramètre variable

Sélectionner dans l'explorateur de sources une des URL parmi celles qui sont analogues puis faire un clic droit : « Filtre avancé ».



Ensuite, cocher la case « Filtre avec un paramètre variable » puis cliquer sur le bouton « Générer » devenu accessible.

L'assistant parcourt l'ensemble des URL de la source et les compare à l'URL sélectionnée pour détecter le paramètre variable et proposer un filtre.
Il convient ici, comme pour les filtres standards, de définir son niveau d'application et son type.

C'est la méthode la plus conviviale pour définir un filtre à paramètre variable mais elle exige que l'on ait effectué un crawl avec un niveau de profondeur suffisant pour atteindre les URL dites « analogues ».
Il est aussi possible de déterminer le filtre à paramètre variable soi-même et l'insérer directement dans les filtres de la source (onglet « filtre »).

3.7.3.6 Filtre de type « Pattern »

Un filtre de type « pattern » s'applique aux URL qui contiennent une chaîne de caractère spécifique (Pattern). Ainsi, on peut « black-lister » ou rendre exclusives des URL qui contiennent ce pattern.

Le pattern n'a pas besoin d'être placé entre crochets.

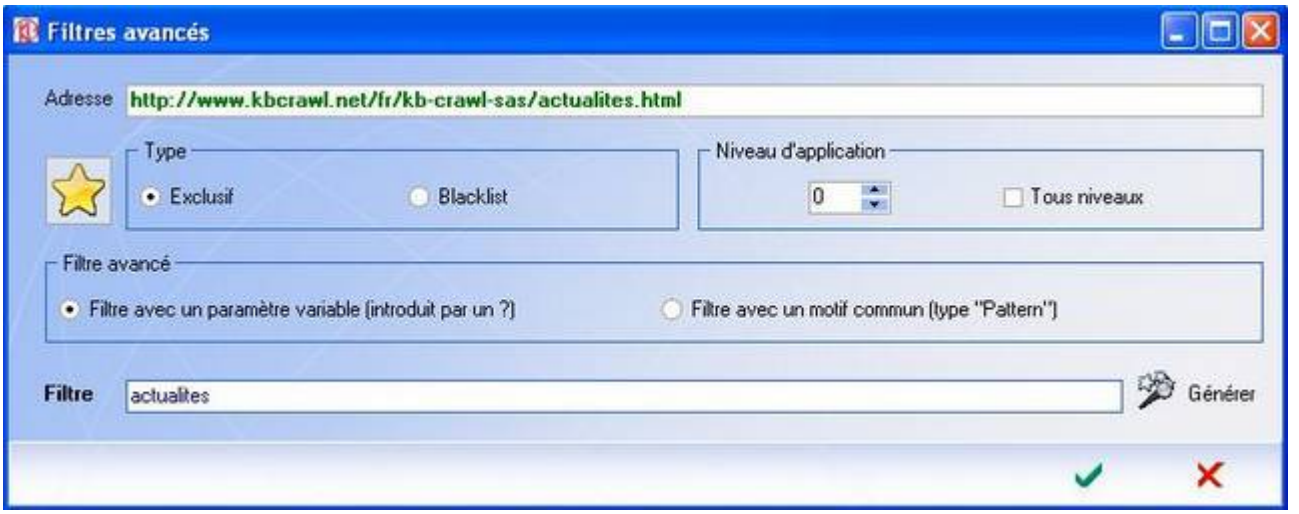
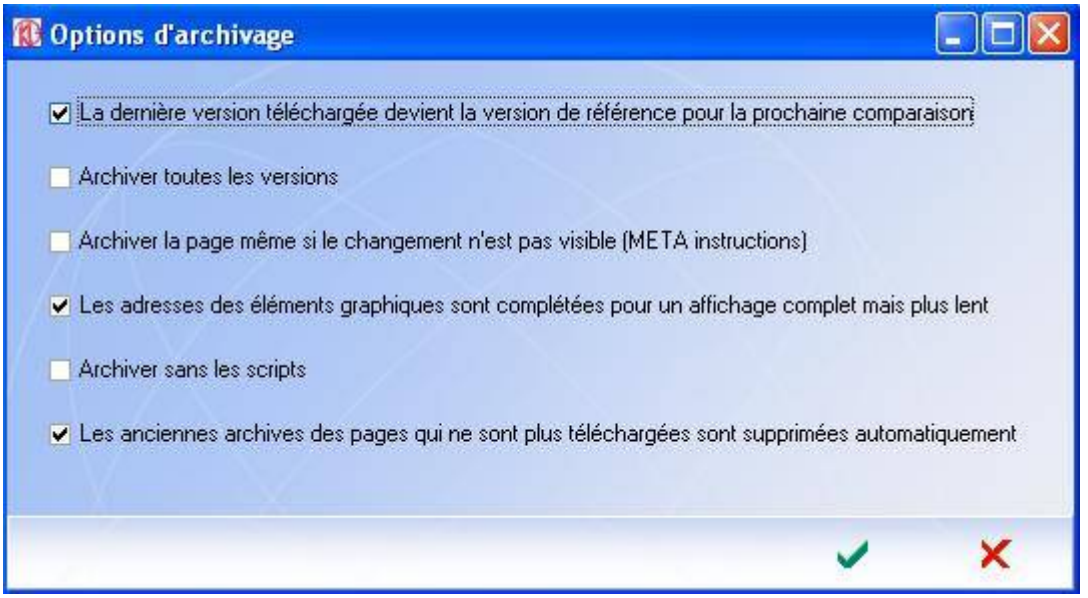


Figure 23 : Filtre de type "Pattern"

Dans l'exemple ci-dessus, un filtre de type « Pattern » est défini pour que seules les URL qui contiennent la chaîne de caractère « actualités » soient récupérées lors du crawl.
Si on coche « black-liste » à la place d' « Exclusif », toutes les URL contenant la chaîne de caractère « actualités » seront black-listées.

3.7.4 Archives



- La dernière version téléchargée devient la version de référence pour la prochaine comparaison

Cette option permet de définir si la version de référence de chaque page demeure celle stockée lors du tout premier crawl ou si celle-ci est mise à jour avec la version 2, c'est à dire l'avant dernière version observée de cette page.

Exemple :
A une date « t », KB Crawl analyse et stocke une page : P1
A une date « t2 », KB Crawl analyse et stocke une nouvelle version de cette page : P2
A une date « t3 », KB Crawl analyse une nouvelle version de cette page : P3.

Si l'option « mise à jour automatique » est sélectionnée, P1 devient P2 et P3 devient P2.
Dans le cas contraire, P2 devient P3 alors que P1 reste la page de référence : Si KB Crawl analyse de nouveau la page au stade Pn, celle-ci sera à chaque fois comparée à P1.

Si on a choisi l'option « Archivage des différentes versions », les pages Pn qui auraient dû être écrasées sont stockées et restent consultables à tout moment depuis le module de gestion d'archives.

- Archiver toutes les versions

Comme vu dans le chapitre 1.7 (fonctions d'archivage), KB Crawl permet d'archiver dans sa base de données chaque page qu'il visite, en plus de la version de référence et de la dernière version observée d'une page.
Choisir cette option indique à KB Crawl qu'il doit stocker une page à chaque fois que celle-ci présente un changement perceptible.

- Archiver la page même si le changement n'est pas visible (META instructions)

Pour archiver la page quel que soit le type de changement survenu (cela peut être utile notamment lorsque des images ont changé), cocher la case «Archiver la page même si le changement n'est pas visible (META instructions) ».

- Les adresses des éléments graphiques sont complétées pour un affichage complet mais plus lent

Une balise de redirection de tous les liens relatifs est insérée dans le code source de chaque page archivée, ce qui permet de visualiser dans ces pages les images ou les frames (= cadres). Dans certains cas, cette redirection empêche de visualiser la page et il convient donc de ne pas choisir cette option en décochant la case « redirection des liens ».

- Archiver sans les scripts

Certaines pages contenant du code JavaScript posent des problèmes d'affichage lorsque l'on tente de les visualiser hors contexte; généralement, l'affichage de la page se fait attendre pour parfois même ne rien obtenir.

- Les anciennes archives des pages qui ne sont plus téléchargées sont supprimées automatiquement

Lorsqu'un document est identifié comme supprimé, le comportement par défaut du module de gestion des archives est de supprimer automatiquement ce document, afin d'optimiser la place occupée par la base d'archive.
Ce comportement par défaut peut être modifié afin de conserver dans la base de données les anciennes versions des documents qui ont été marqués comme supprimés ; pour cela, il suffit de décocher la case « Les anciennes archives des pages qui ne sont plus téléchargées sont supprimées automatiquement ».

3.7.5 Paramètres avancés

- Variables de sessions

Beaucoup d'URL que l'on rencontre sur le Web comportent des paramètres.

Exemple, URL n°1 :
http://www.openlaszlo-france.com/phpBB2/viewforum.php?f=1&sid=58522fcbc5967bf59cc4d11b74a26e

A partir du caractère « ? », on trouve une série d'expressions de type « paramètre=valeur » qui sont toutes séparées par le caractère « & ».
Ici, on a, entre autres :
F=1

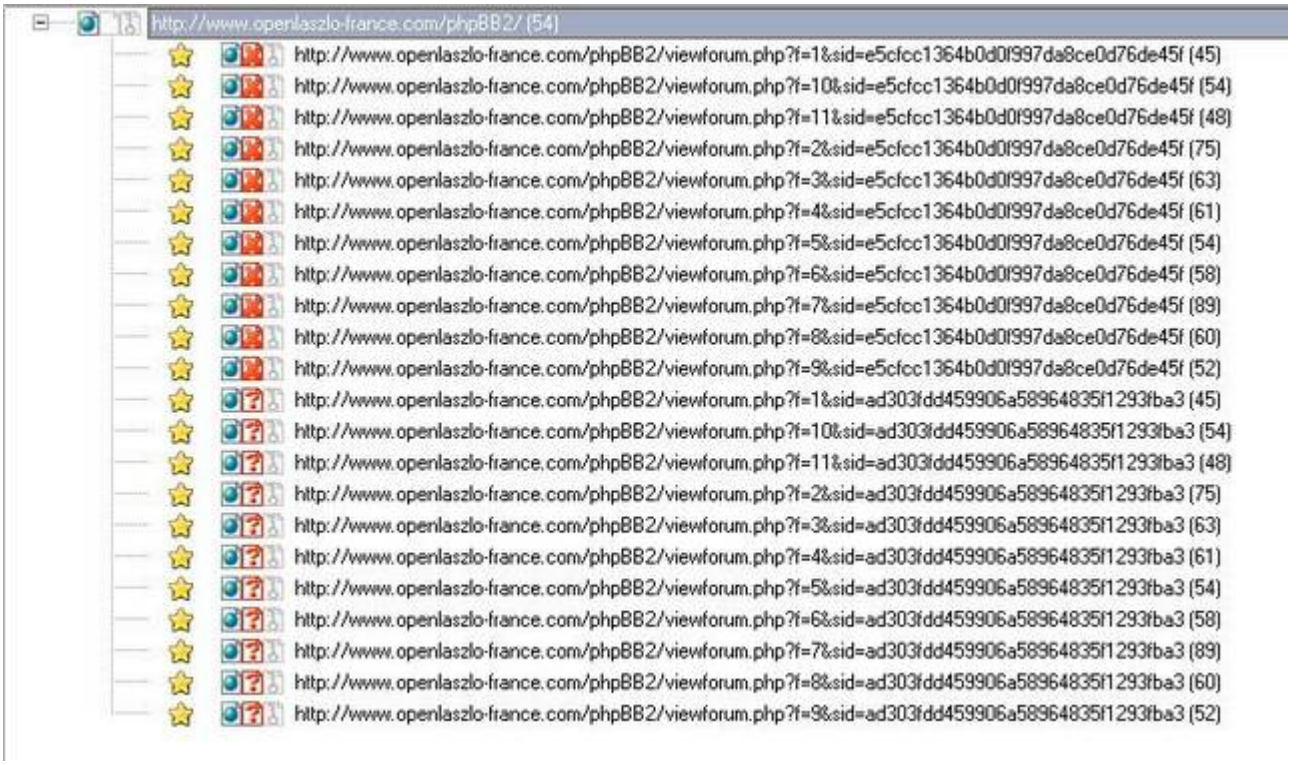
sid=58522fcbc5967bf59cc4d11b74a26e
Ces paramètres sont des informations qu'interprète le serveur lorsqu'il reçoit la requête HTTP du client.
L'URL mentionnée ici a été trouvée sur le lien d'une page Web. A la prochaine session* ouverte avec le serveur, le même lien (qui mènera sur la même page que la fois précédente) recouvrera une URL légèrement différente.

*Lorsque l'on ouvre un navigateur par exemple, on ouvre une nouvelle session et on ferme cette session lorsque l'on referme ce même navigateur.

Exemple, URL n°2 :
http://www.openlaszlo-france.com/phpBB2/viewforum.php?f=1&sid=649f0ef4894c807dc77f71c9a19fb5b0

En effet, un paramètre a changé : SID.
Sa valeur est différente par rapport à la session précédente et le changement de sa valeur ne change en rien la page qui correspond à cette URL : on appelle cela une variable de session.
A chaque crawl, KB Crawl ouvre une session différente, si bien que lors d'un crawl de comparaison (3.8), cette URL sera considérée comme nouvelle. Elle l'est, si on considère la chaîne de caractère qui constitue l'intégralité de l'URL, mais la page Web qui y correspond n'est, en fait, pas nouvelle.
Pour éviter, lors d'un crawl de comparaison, que l'URL n°1 soit considérée comme supprimée et l'URL n°2 comme nouvelle, il convient d'ignorer ce paramètre en le spécifiant dans le champ « paramètres ignorés pour chaque URL ». Il peut y avoir plusieurs paramètres à ignorer, dans ce cas, ils doivent être séparés par des « ; ».

L'exemple ci-dessous montre ce qui se passe lorsque l'on effectue un crawl de comparaison sur un site qui comporte des variables de sessions.
Il semble, d'après l'explorateur, que beaucoup de pages ont été supprimées et autant ajoutées. La coïncidence laisse supposer la présence de variables de session ou encore, paramètres à ignorer.



L'exemple ci-dessous montre ce qui se passe lorsque l'on effectue un crawl de comparaison après avoir saisi « sid » comme paramètre à ignorer :

- Authentification (accès à un espace sécurisé)

Certaines URL sont protégées par ce que l'on appelle une authentification de base : le serveur Web qui les héberge exige de la requête HTTP qui demande leur téléchargement de présenter les paramètres d'authentification requis.
Lorsque l'on cherche à télécharger un page protégée par une authentification de base, on voit surgir une boîte de dialogue depuis le navigateur que l'on utilise :

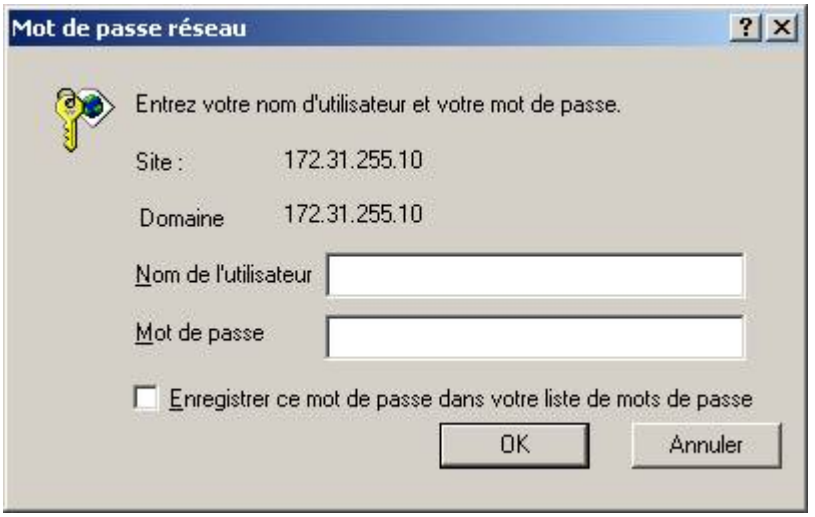


Figure 24 : Boîte de dialogue d'une authentification de base.

Cette boîte de dialogue n'apparaît pas lorsque KB Crawl cherche à télécharger une page Web parce qu'il envoie directement sa requête au serveur et reçoit sa réponse sans passer par l'interface d'un navigateur.
KB Crawl intègre les paramètres d'authentification de base une fois pour toutes au sein du paramétrage d'une source.
Il faut toutefois vérifier au préalable que le téléchargement requiert effectivement cette authentification à l'aide d'un navigateur Web. Si c'est le cas, cocher la case située devant « Authentification (accès à un espace sécurisé)» puis saisir le nom d'utilisateur et le mot de passe requis.

- Parties d'URL à ignorer

Il apparait sur certains sites des adresses URLS qui sont différentes à chaque connexion, même si la page résultante est la même.
Afin de s'affranchir d'alerte sur pages disparues et/ou nouvelles, il est possible de masquer certaines parties de ces URLS.
Ainsi, il faut saisir une borne inférieure et une borne supérieure qui sont des chaines de caractères statiques contenues dans les adresses URL à traiter. La borne supérieure peut être vide, signifiant ainsi que la limite sera la fin de l'adresse URL.

- Options de sécurité

- User Agent

Lorsqu'une requête HTTP est émise, l'entête de celle-ci, qui est non visible pour l'utilisateur, contient une chaîne de caractères qui représente l'identité du logiciel qui a émis la requête HTTP, c'est ce que l'on appelle le « User Agent ».
Certains serveurs exigeants refuseront de répondre à une requête si celle-ci ne présente pas un « User Agent » qu'il considère comme valide.
Par défaut, ce paramètre est initialisé avec un « User Agent » qui correspond à Internet Explorer version 6.0.
Dans certains cas particuliers, il peut être nécessaire de saisir un « User Agent » spécifique.

- Utiliser la configuration de sécurité d'Internet Explorer

Cette option est utile dans certains cas très spécifiques, notamment pour crawler des sites ayant un niveau de sécurité avancé. En activant cette option, le crawl de la source s'appuiera sur Internet Explorer, il est donc indispensable de disposer de la version 6 d'Internet Explorer au minimum, et de le configurer correctement afin qu'Internet Explorer ait accès à internet (proxy, etc.).

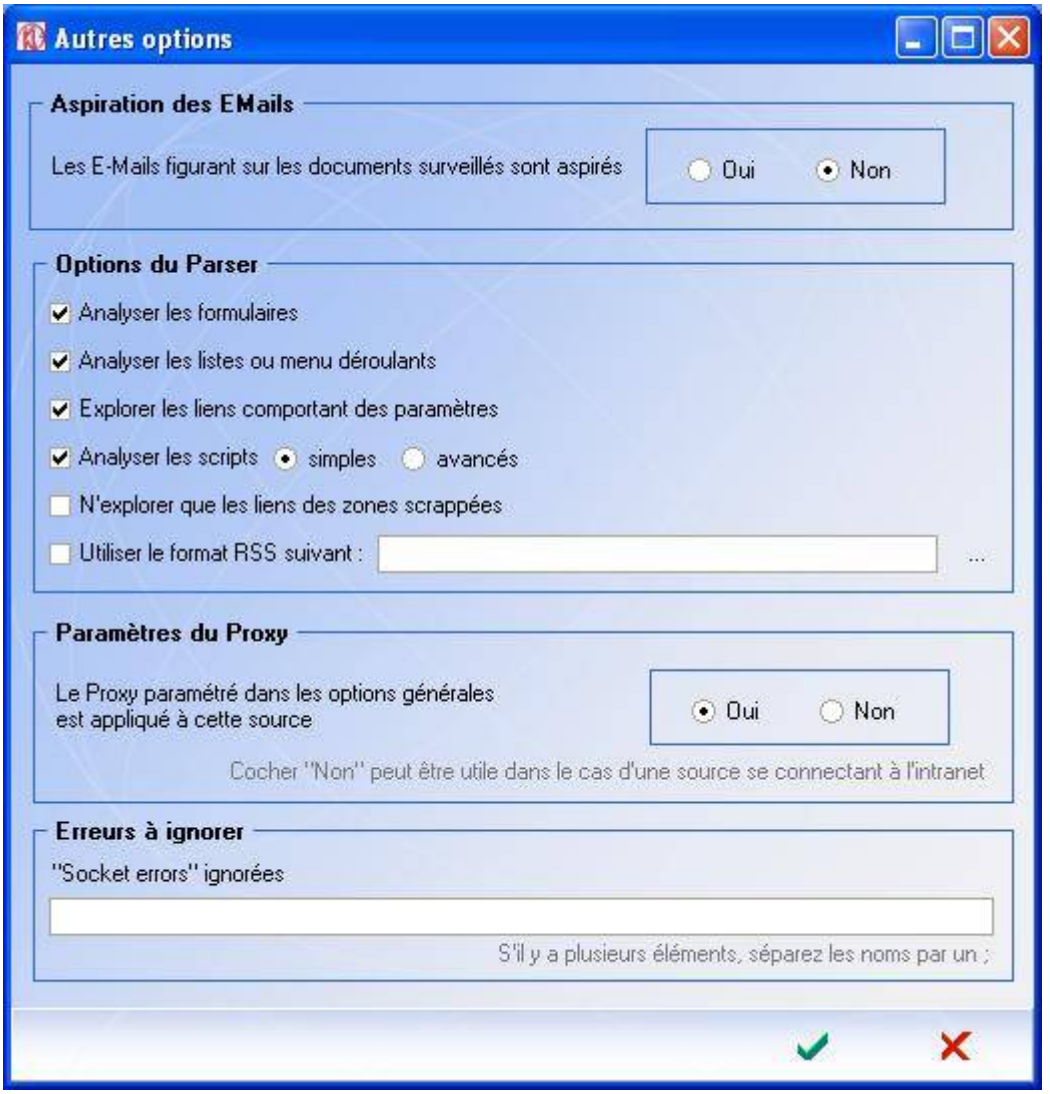
- Prise en compte du fichier ROBOTS.TXT

Le fichier ROBOTS.TXT se trouve à la racine de certains sites et s'adresse aux robots de type KB Crawl qui sont amenés à télécharger un certain nombre de pages de ce site. Il mentionne pour chaque robot (ou pour tous les robots) la liste des chemins et documents pour lesquels le téléchargement leur est « interdit ».
Pour que KB Crawl analyse ce fichier avant chaque crawl et tienne compte des interdictions qui y sont inscrites, cocher l'option « prendre en compte le fichier ROBOTS.TXT ».

- Temporisateur

Une fois activé (en cochant la case « Activer le temporisateur »), une pause est marquée entre chaque téléchargement. La durée de cette pause est comprise entre les deux valeurs saisies (en ms).

3.7.6 Autres options



3.7.6.1 Aspiration des E-Mails

Au cours de son exploration, KB Crawl peut rencontrer, à l'intérieur des pages qu'il analyse, des adresses e-mail. Par défaut, il les ignore. En cochant l'option « Les E-mails figurant sur les documents surveillés sont aspirés », toutes les adresses e-mail figurant sur les pages crawlées seront collectées et placées dans le menu « Affichage > E-mail » : « Liste des e-mails aspirés ».

3.7.6.2 Options du Parser

- Analyser les formulaires
  Les formulaires que l'on rencontre sur les pages Web contiennent des adresses qui sont visitées lorsque l'on valide le formulaire (« envoyer »).
  On a le choix ici de visiter ces adresses ou non.

- Analyser les listes ou menus déroulants
  Certaines pages Web contiennent des listes déroulantes dites « navigantes » :
  Lorsque l'on sélectionne un élément qu'elles proposent, on navigue vers une autre page. Parfois, le code HTML qui produit ces listes déroulantes contient explicitement ces liens.
  Lorsque l'on choisit cette option, KB Crawl va reconstruire un lien à partir de la valeur qu'il trouve derrière chaque élément de ces listes déroulantes et tenter de visiter ce lien.
  Ne pas cocher cette option permet d'optimiser le temps d'exploration en évitant de visiter des liens inutiles quand ces listes déroulantes ne sont pas navigantes.

- Explorer les liens comportant des paramètres
  Ces liens sont fréquemment appelés « fat URL » en anglais. Ceci permet au serveur, lorsqu'il reçoit la requête HTTP du poste client de construire dynamiquement la page, qui constitue sa réponse, en fonction de ces paramètres.

  Exemple :
  http://www.monsite.com/search/DetailArticle?PK=420&Order=DATECREATION+desc&StartRec=1&StartPageLink=1&Filter=&SID=478692269&DetailMode=Modif

  Le premier paramètre est toujours précédé d'un « ? » et les suivants d'un « & ».
  Ici, le paramètre PK vaut 420 et SID vaut 478692269.
  Les sites dits « dynamiques » étant très répandus, ces types de liens le sont également.

  Ils ne sont pas un obstacle au processus de crawl. Cependant, ils ne sont, dans certains cas, d'aucun intérêt au regard de l'information qu'ils peuvent apporter. Ils peuvent même parfois provoquer ce que l'on appelle une « boucle de crawl ».

  Exemple : Un lien mène à une page construite dynamiquement avec le paramètre suivant PARAMEXEMPLE=1. La page construite dynamiquement présente le même lien avec PARAMEXEMPLE=2 qui va mener vers la même page contenant le même lien avec PARAMEXEMPLE=3 et ainsi de suite…

Si on décoche cette option, KB Crawl ne visitera pas les liens dynamiques.

- Analyse des scripts
  KB Crawl a la capacité d'analyser les scripts « JavaScript » contenus dans le code HTML des pages Web afin d'en extraire les liens qui mènent vers d'autres pages.
  Cette profondeur d'analyse durant la méthode de parsing a un léger coût en termes de performance qui peut être économisé si la source ne visite que des pages ne contenant pas de scripts ou bien si ces scripts ne contiennent aucuns liens.
  On peut donc optimiser le crawl en ne sélectionnant pas cette option.

- N'explorer que les liens des zones scrapées
  Lors de l'export des données par KB Crawl, il peut être utile, dans le cas où l'on utilise KB Scraper, de n'exporter que les zones scrapées des pages concernées ; cette opération est possible grâce à cette option.

- Utiliser le format RSS suivant
  Si la source contient un flux RSS, la transformation RSS vers la page HTML se fera avec le fichier par défaut (FormatRSS.xsl). Il est possible avec cette option de modifier le fichier XSLT pour une seule source en particulier.

3.7.6.3 Paramètres du Proxy

Dans certains cas, notamment lorsque l'on veut crawler des documents situés sur un intranet, les paramètres de connexion à Internet via un serveur Proxy ne doivent pas s'appliquer, il faut alors cocher « non ».

3.7.6.4 Erreurs à ignorer

Lors du téléchargement, KB Crawl peur rencontrer ce que l'on appelle en terme technique des « socket errors », le téléchargement du document est alors annulé.

On peut changer ce comportement en listant les numéros de « socket errors » que l'on veut ignorer.

Il suffit alors de placer entre « ; » les numéros des « socket errors » en question dans le champ prévu à cet effet.

Par exemple, lors du crawl d'une source, si la boite de dialogue de KB Crawl affiche le message : « http://www.monsite.com : échec au téléchargement (HTTP/1.1 404 Not Found) Lien trouvé à l'URL : http://www.monsite.com » il est possible d'ignorer cette erreur en spécifiant dans le détail de la source, erreur à ignorer : « HTTP/1.1 404 Not Found ». Ainsi, le message d'erreur ne s'affichera plus pour cette source.

3.7.7 Commentaires

Le menu commentaire ouvre une boîte de dialogue qui permet d'enregistrer des commentaires liés à la source.

3.7.8 Lancement du crawl

Lorsqu'une source a été créée, et qu'elle figure dans la liste des sources, elle est prête à être explorée (ou crawlée) par KB Crawl.

Pour lancer un premier crawl, il faut cliquer sur le bouton « Crawl »  situé sur la barre d'outils générale.

- La barre de progression donne une indication sur les liens issus de la page de départ qui sont parcourus.

  Exemple : KB Crawl analyse la page départ spécifiée dans la source et trouve 10 liens : au 6ème lien, la barre de progression est à 50%.
  Cette progression ne peut pas donner d'avantage d'informations sur le temps restant pour achever l'exploration, tout dépend de la « profondeur » de chacun de ces liens rencontrés sur la page de départ :
  Les premiers liens ont peut-être 3 ou 4 pages filles alors que le 6ème en a 500 !
  Dans ce cas, la barre de progression va vite arriver à 50 % pour y rester le temps de crawler les 500 pages du lien n°6.

- La barre d'état juste en-dessous de la barre de progression indique l'action globale que KB Crawl est en train de réaliser, du résultat une fois que le traitement est terminé, ou d'un message d'erreur éventuel (dans ce cas, le message apparaît en rouge).

- La barre d'état en bas à gauche de la fenêtre informe du nombre de pages crawlées avec succès : celles pour lesquelles le crawl a été jusqu'au stockage du contenu dans la base de données.

- La barre d'état en bas à droite indique quelle est l'étape en cours durant le crawl d'une page. Les libellés que l'on peut voir apparaître sont successivement :

  - Téléchargement de la page : *adresse complète de la page*
  - Extraction du contenu pour : *adresse complète de la page*
  - Terminé en : *temps total mis pour l'ensemble de la source*

3.8 Comparaison

3.8.1 Le processus de comparaison (fonctionnement)

Pour contenir les données relatives à une source, la base de données de KB Crawl possède deux espaces de stockage distincts.
On les nommera espaces de stockage n°1 et n°2.

L'espace n°1 sert à stocker tout le contenu d'une source lors d'un premier crawl. Ce sont les URL contenues dans cet espace que l'on voit dans l'explorateur.

L'espace n°2 sert à stocker la dernière version des pages téléchargées, si toutefois la dernière version présente des changements par rapport à la précédente version.

Lors d'un crawl de comparaison, dès qu'une page a été téléchargée et que le contenu textuel en a été extrait, KB Crawl recherche la page correspondante.
On appellera page P1bis la page qui vient d'être téléchargée et page P1 la page contenue dans l'espace n°1 qui lui correspond parce que leurs adresses sont identiques.

KB Crawl compare ces deux pages et observe les cas suivants :

- Il n'y a pas de page P1 dans l'espace n°2 :

P1bis est une page nouvelle. Chaque nouvelle page identifiée est insérée temporairement dans l'espace n°1, afin de visualiser sa présence, et dans l'explorateur de sources (4).

- Le nombre de liens contenus dans les pages P1 et P1bis est différent :

Si la valeur absolue de la différence entre ces deux nombres dépasse le seuil d'alerte défini dans les options de la source, la page 1 est marquée comme ayant son nombre de liens changé (le nombre de ces liens est stocké dans les espaces n°1 et n°2).

- Des mots-clés d'alerte sont apparus :

KB Crawl a en mémoire au moment de la comparaison le nombre d'occurrences de chaque mot-clé d'alerte trouvé dans la page P1. Si dans la page P1bis l'un de ces

mots-clés référencés apparait plus de fois que dans la page P1, la page P1 est marquée comme ayant de nouveaux mots-clés apparus.

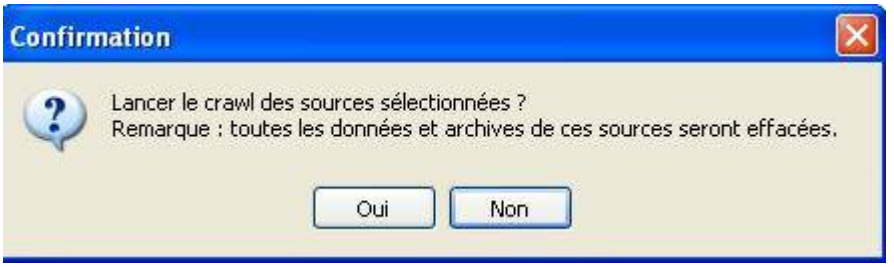
- Les contenus textuels des pages P1 et P1bis sont différents :

KB Crawl compte alors le total de mots de chacune des deux pages et observe si la valeur absolue de la différence entre les deux totaux dépasse le seuil d'alerte défini dans les options. Si oui, la page 1 est marquée comme ayant son contenu textuel changé.

3.8.2 Lancer la comparaison

Pour cela, il est nécessaire de se positionner sur une source dans l'arborescence et de cliquer sur le bouton « Comparer »  situé sur la barre d'outils générale.
Les effets au niveau de l'interface sont alors les mêmes que ceux observés durant le premier crawl.

Important : Une source qui a déjà fait l'objet d'un premier crawl peut être de nouveau crawlée (crawl d'initialisation). Cela revient à vider les espaces n°1 et n°2 des contenus de la source qu'il contient et à tout reprendre depuis le début.
Dans ce cas, une boite de dialogue nous invite à confirmer l'écrasement des données :



4 L'explorateur de sources

4.1 Généralités

Une source contient au minimum une page Web.
Dans le cas où elle représente un ensemble de pages, celle-ci est hiérarchisée : la page de départ a fourni un certain nombre de liens qui ont mené vers des pages qui elles-mêmes fournissent d'autres liens, etc.
Cet ensemble de pages, qui représente une partie ou l'intégralité d'un site Internet, peut être représenté sous forme d'un arbre afin d'en observer sa structure.
C'est ce que fait l'explorateur de sources.

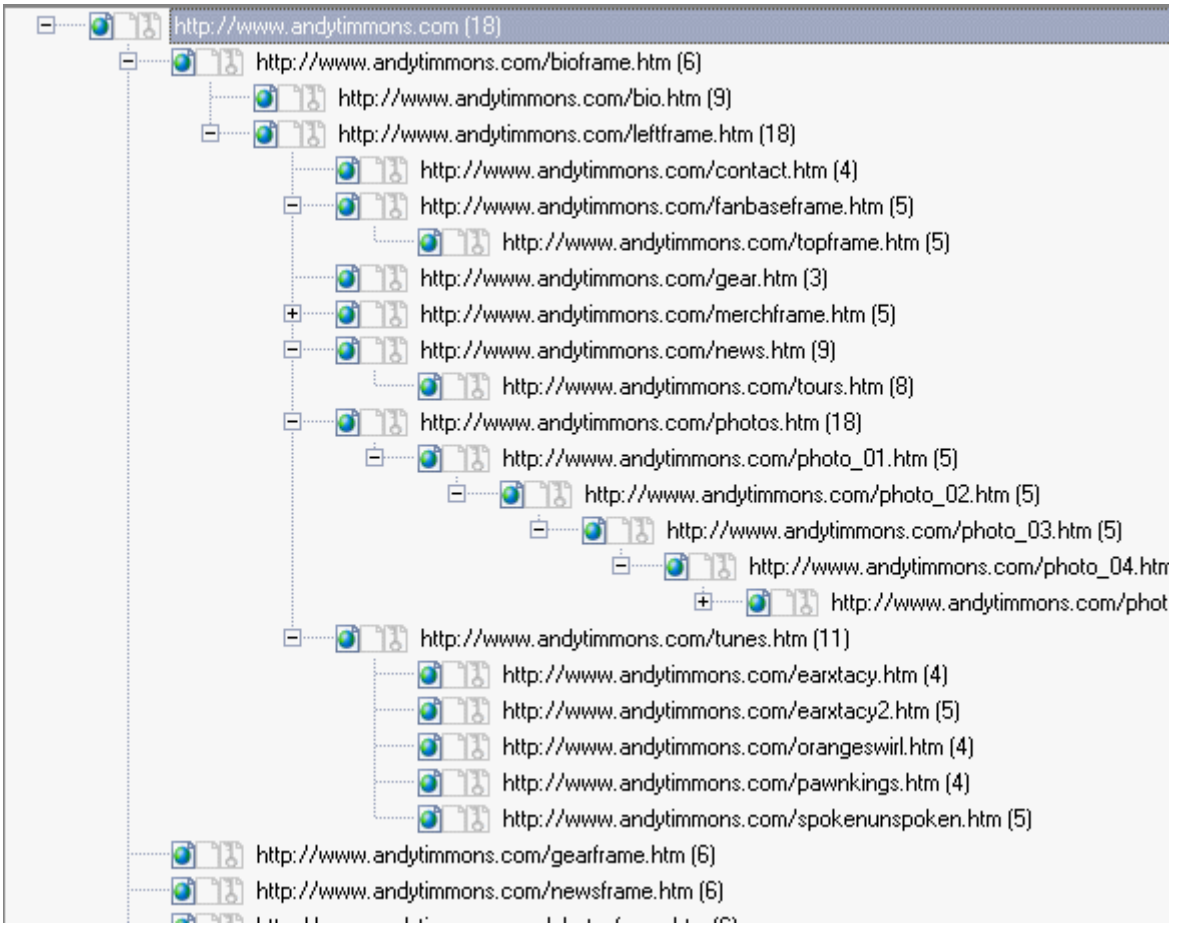


Figure 25 : L'explorateur de sources

Ici, la source indique une adresse de départ qui est http://www.andytimmons.com avec une profondeur de page infinie et une profondeur de site qui vaut 0.
Chaque nœud de l'arbre représente donc une page du site.
Le libellé contenu dans chacun d'eux représente l'URL complète de la page et le nombre entre parenthèses correspond au nombre de liens qui ont été extraits de cette page (qu'ils aient déjà été visités ou non) et dont l'extension a été définie dans la bibliothèque d'extensions.
La structure du site qui apparaît représente la partie essentielle du plan du site Internet.
Le lien de parenté entre deux pages dépend de l'ordre dans lequel les liens du site ont été visités, et un lien n'est visité qu'une seule fois lors d'une exploration.

4.2 Utilisation et ergonomie générale

Lorsque l'on ouvre une source avec l'explorateur, l'arbre est déplié par défaut. On peut ouvrir et fermer chacun des nœuds en cliquant dessus (sur la croix) et faire ainsi apparaître ou disparaître les pages filles du nœud sélectionné.

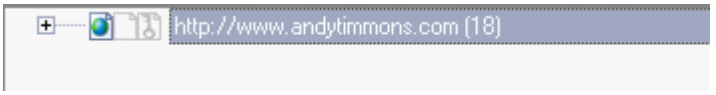


Figure 26 : Arbre replié dans l'explorateur de sources.

On peut aussi dérouler et replier l'arbre entièrement par un clic droit / Dérouler entièrement (Ctrl-D) ou le replier entièrement par un clic droit / Replier entièrement (Ctrl-Alt-D).

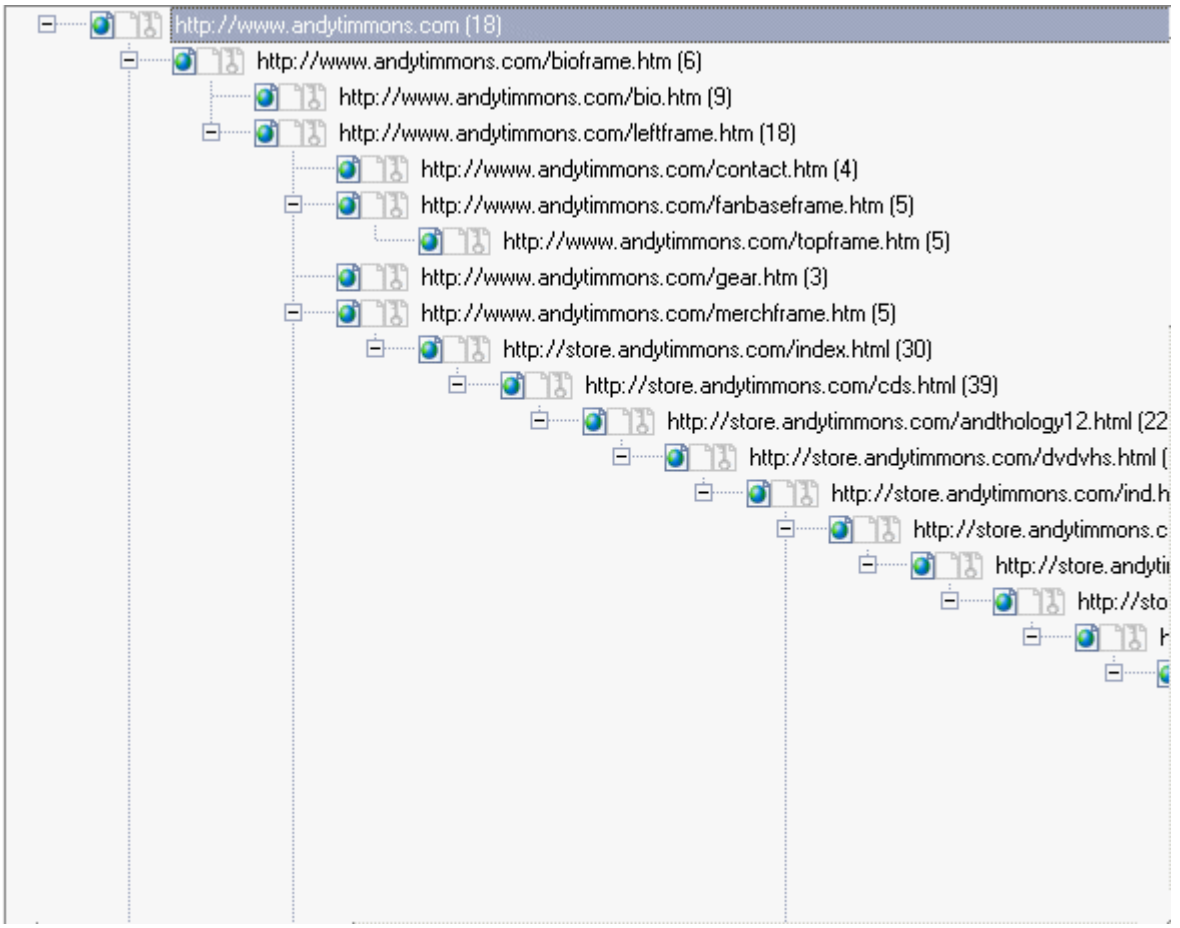


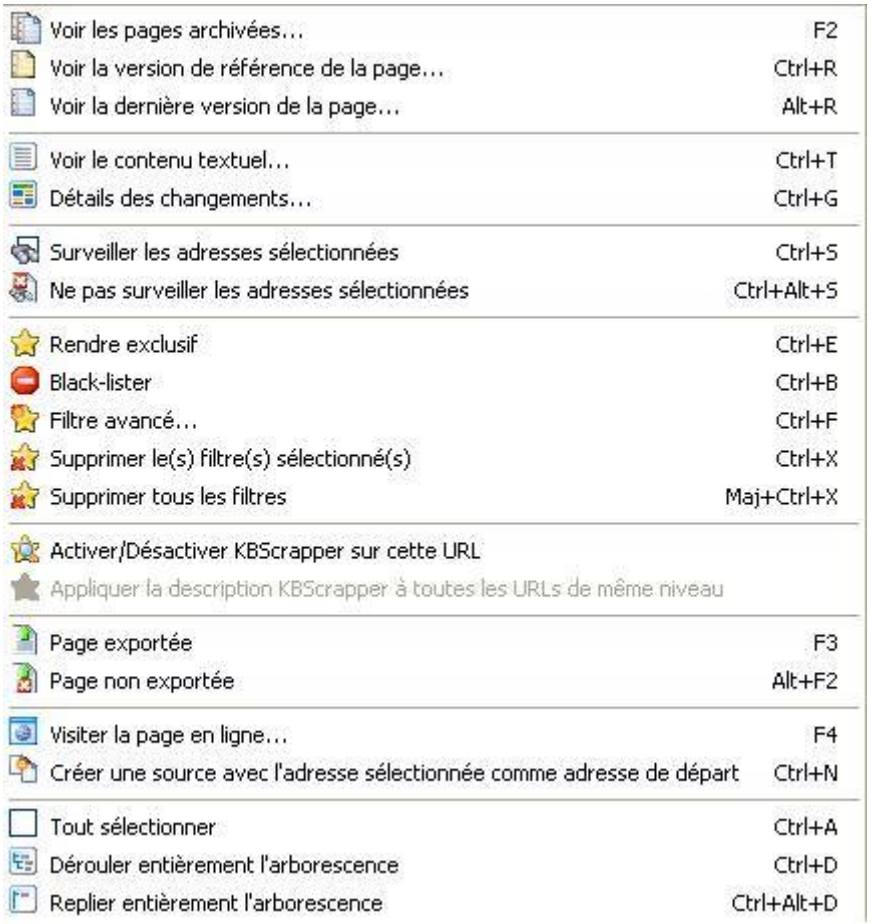
Figure 27 : Arbre entièrement déplié dans l'explorateur de sources.

4.3 Fonctionnalités à partir de l'arbre

L'arbre est constitué de nœuds représentant des URL qui ont été crawlées par KB Crawl. Une surbrillance grise indique qu'un nœud (ou URL) est sélectionné.

Figure 28 : Sélection d'une URL dans l'explorateur de sources.

Toutes les fonctionnalités de l'explorateur de sources sont accessibles à partir d'une URL (ou nœud) de l'arbre par un clic droit qui fait apparaître un menu contextuel :



Toutes ces fonctionnalités sont donc contextuelles à une URL.

4.3.1 Voir les pages archivées

Se reporter au chapitre « Gestionnaire d'archives » (§ 5)

4.3.2 Voir la version de référence de la page

Affiche la version de référence de la page sélectionnée dans l'onglet « Browser » §.1.7.2

4.3.3 Voir la dernière version de la page

Affiche la dernière version de la page sélectionnée dans l'onglet « Browser » §.1.7.2

4.3.4 Voir le contenu textuel

Le contenu textuel de chacune des pages correspondant aux nœuds de l'arbre est enregistré dans la base de données de KB Crawl et est consultable de la façon suivante :
Se positionner sur n'importe quel nœud puis faire un clic droit / Voir le contenu textuel (Ctrl-T).

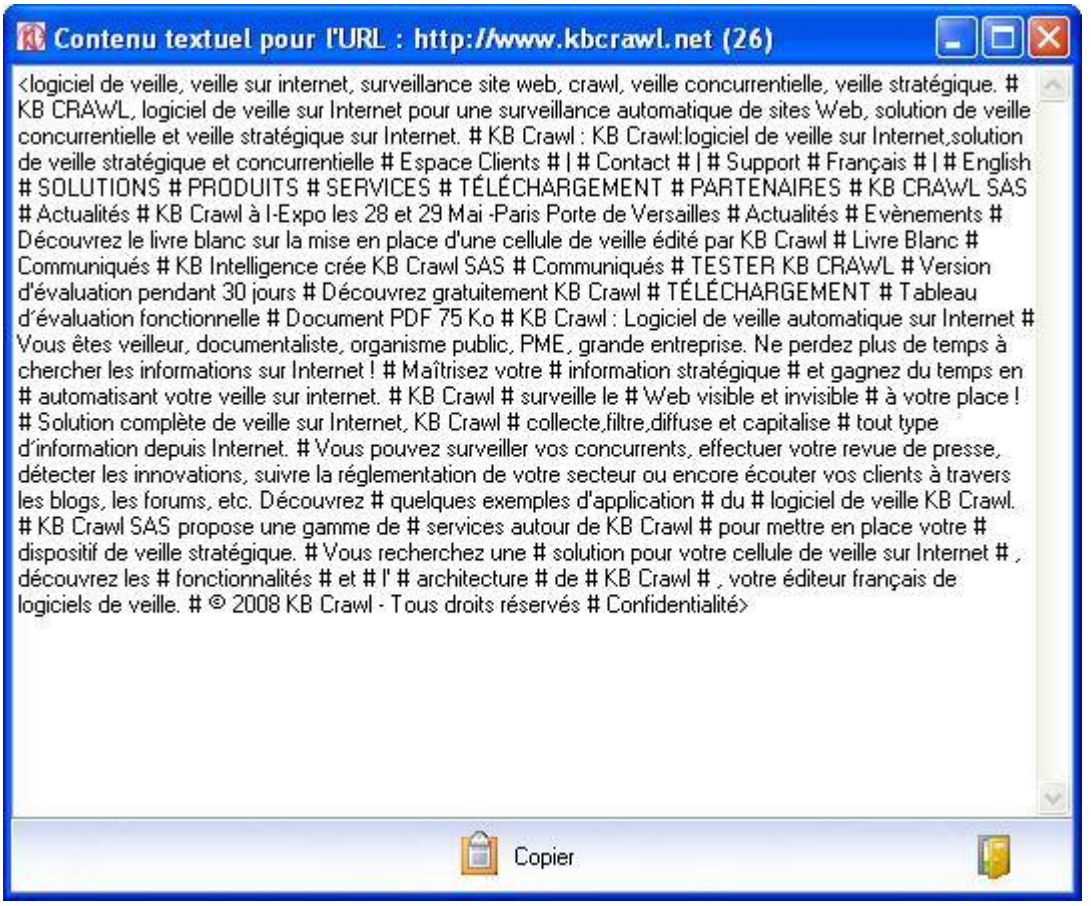


Figure 29 : Contenu textuel d'une URL.

Remarque : Les # sont des séparateurs de blocs de textes et deux # devant un mot indiquent un lien.

4.3.5 Détails des changements

Lorsque l'on est positionné sur une page de l'arbre, on peut prendre connaissance des détails des éventuels changements survenus sur cette page en faisant apparaître une fenêtre qui montre un comparatif entre la première version de la source (espace n°1) et celle issue d'un crawl de comparaison (espace n°2). Ce tableau comparatif met en évidence les éléments qui font l'objet d'un changement : nombre de mots dans la page, nombre de liens et liste des mots-clés qui font l'objet d'une surveillance.

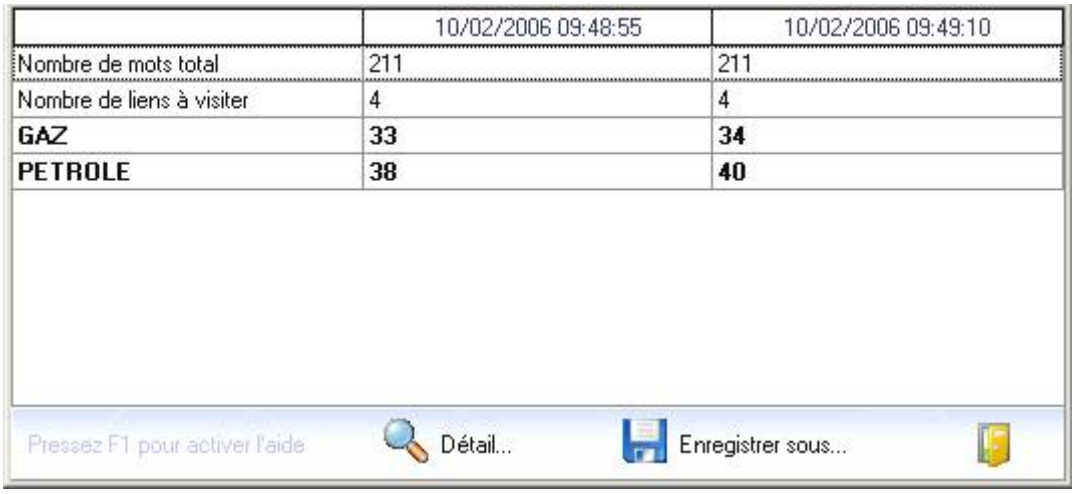
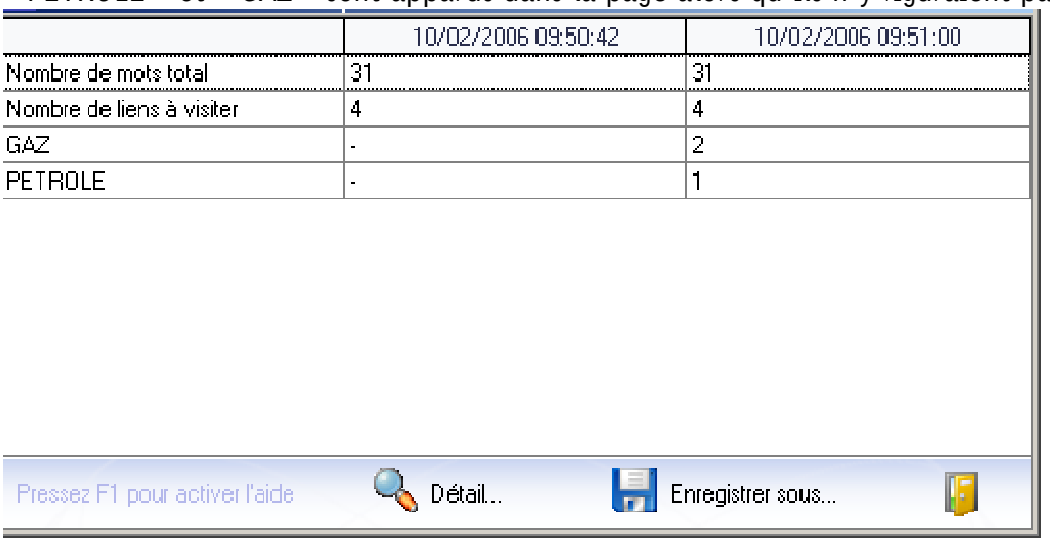


Figure 30 : Détail des changements pour une URL.

Le tableau contenu dans cette fenêtre montre un comparatif entre la page à la date et l'heure où elle a été intégrée dans l'espace de stockage n°1 et cette même page au moment de son intégration dans l'espace de stockage n°2, lors du dernier crawl de comparaison.
Dans le cas observé, on voit que le nombre total de mots a augmenté et que les mots-clés « PETROLE » et « GAZ » sont apparus dans la page alors qu'ils n'y figuraient pas auparavant.



KB Crawl permet de localiser plus précisément ces mots-clés d'alerte sur la page archivée.
En cliquant sur le bouton « Détail… », un navigateur s'ouvre pour visualiser la page Web qui présente les changements détectés lors du processus de comparaison.

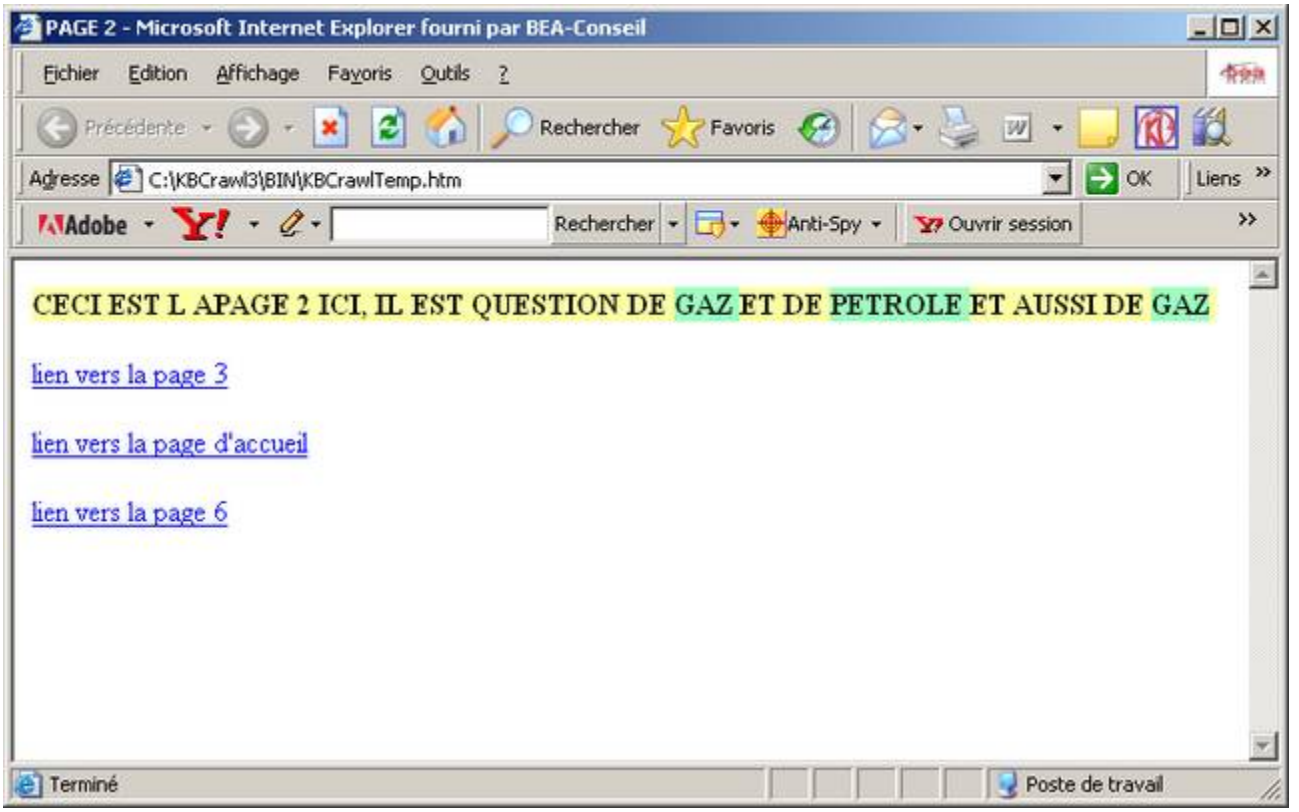


Figure 31 : Surlignement des changements dans une page.

Si l'alerte porte seulement sur un changement de contenu et non sur l'apparition de mots-clés d'alerte, ce sont les blocs de textes qui ont changé et qui sont surlignés :

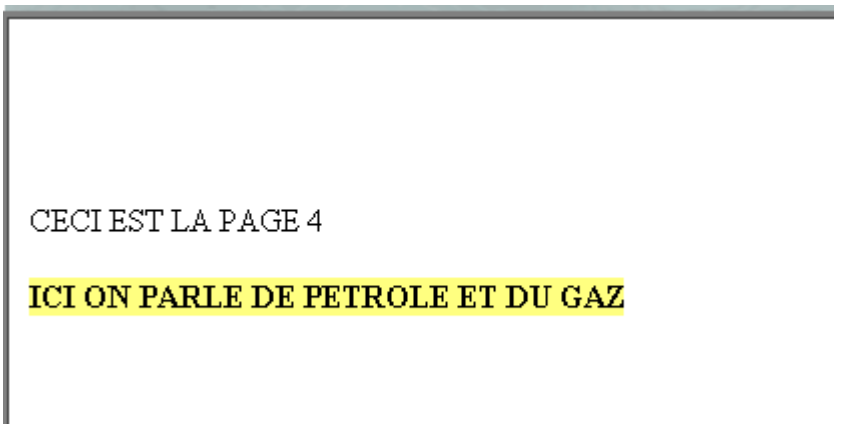


Figure 32 : Surlignement des changements apparus dans une page.

On peut également exporter le tableau comparatif dans fichier Excel en cliquant sur le bouton « Enregistrer sous… ».

4.3.6 Rendre Exclusif

Rend l'URL sélectionnée de l'arbre exclusive (§.3.7.3.1)

4.3.7 Black-lister

Black-liste l'URL sélectionnée dans l'arbre. (§ 3.7.3.2)

4.3.8 Supprimer le(s) filtre(s) sélectionné(s)

Supprime tous les filtres appliqués aux URL sélectionnées

4.3.9 Supprimer tous les filtres

Supprime tous les filtres d'une source sélectionnée.

4.3.10 Filtre avancé

Appelle l'assistant pour les filtres avancés.

4.3.11 Activer/Désactiver KB Scraper sur cette URL

Permet d'accéder aux menus de KB Scraper. Pour connaître le fonctionnement de KB Scraper, se reporter au manuel utilisateur KB Scraper.

4.3.12 Appliquer la description KB Scraper à toutes les URL de même niveau

Pour connaître le fonctionnement de KB Scraper, se reporter au manuel utilisateur KB Scraper.

4.3.13 Page exportée

Marque la page comme étant à exporter lors de l'export automatique qui a lieu à la fin de la comparaison.

4.3.14 Page non exportée

Marque la page comme n'étant pas à exporter lors de l'export automatique qui a lieu à la fin de la comparaison.

4.3.15 Visiter la page en ligne

Chaque nœud de l'arbre représente une URL ou adresse Internet qui peut être visitée et visualisée dans le navigateur défini par défaut.
Se positionner sur n'importe quel nœud puis faire un clic droit / Visiter la page en ligne.

4.3.16 Créer une source avec l'adresse sélectionnée comme adresse de départ

Cette fonctionnalité permet de créer une nouvelle source pour laquelle l'adresse de départ est l'adresse d'un nœud quelconque de l'arbre :

Se positionner sur n'importe quel nœud puis faire un clic droit / Créer une source avec l'adresse sélectionnée comme adresse de départ (Ctrl-N).
La source ainsi créée se trouvera dans le même dossier que la source dont elle est issue avec une profondeur de site et de page égale à 0 et une configuration d'options par défaut.
Il suffit ensuite de définir précisément son paramétrage.

Cette fonctionnalité est très utile lorsque l'on veut explorer plus en profondeur une partie d'un site après l'avoir localisée précisément.

4.4 Les différentes icônes de l'arbre

Chaque nœud de l'arbre porte un triptyque d'icônes qui permet de voir d'un coup d'œil :

- le type de document dont il s'agit,
- si le document est en alerte, et de quel type d'alerte il s'agit.

En utilisant le menu « Affichage », on peut voir une légende complète de chaque icône qui participe à la combinaison de trois icônes significatives :

Icone	Catégorie	Commentaire
	Source	Source qui ne comporte aucun changements
	Source	Source qui comporte des changements
	Source	Dossier ou sous-dossier
	Source	Groupe de recherche
	Source	Source verrouillée
	Document	Dossier ou sous-dossier FTP
	Document	Document HTML ou texte
	Document	Formulaire Web
	Document	Document Flash
	Document	Document RSS
	Document	Document Acrobat PDF
	Document	Document Word
	Document	Document Excel
	Document	Document Power Point
	Document	Message de groupe de news
	Document	Document Image
	Alerte	Aucune alerte sur le document
	Alerte	Le contenu du document a changé
	Alerte	Mot(s)-clé(s) trouvé(s) dans le document
	Alerte	Le document est nouveau
	Alerte	Le document a été supprimé
	Filtre	Fitre exclusif
	Filtre	Filtre blacklist
	Filtre	Filtre lien forcé
	Filtre	Page non surveillée

Voici quelques exemples de combinaisons possibles :

- Un document PDF nouveau qui comporte un ou des mots-clés d'alerte :
- Un document HTML ou texte supprimé :
- Un document de type fil RSS dont le contenu a changé et qui comporte un ou des mots-clés d'alerte :

5 Le gestionnaire d'archives

Comme vu dans le chapitre « Fonctions d'archivage » (cf. 1.7.2), KB Crawl est capable d'archiver toutes les versions différentes d'une même page contenues dans une source. Un module spécialement dédié à l'archivage permet de consulter et gérer ces archives.

Figure 33 : Le gestionnaire d'archives.

Lorsque l'on ouvre le module d'archives, celui-ci se positionne automatiquement sur la dernière version archivée de la page.

5.1 Ergonomie générale

Le gestionnaire d'archives, comme la fenêtre générale, adopte une ergonomie de type « explorer » : il est composé de trois parties principales :
L'explorateur d'archives, la barre d'outils générale et le browser d'archives.

5.1.1 L'explorateur d'archives

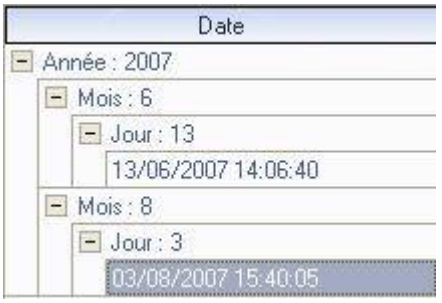


Figure 34 : l'explorateur d'archives

La date d'archivage (année, mois, jour, heure, minute, seconde) de chaque page est enregistrée dans la base d'archives de KB Crawl. Ainsi, toutes les archives d'une même page peuvent être présentées, regroupées et ordonnées par année, mois, jour.
Chaque élément d'archives est présenté sous forme d'un lien cliquable pour visionner le contenu de l'archive dans le browser du module d'archives.
La surbrillance grise montre qu'une archive est sélectionnée.

Un clic droit fait apparaître un menu contextuel qui propose les fonctionnalités suivantes :

Supprimer les archives de l'année
Supprimer les archives du mois
Supprimer les archives du jour
Supprimer le fichier archivé.

L'année, le mois ou le jour sont ceux du fichier (archive) sélectionné.

5.1.2 La barre d'outils générale

Figure 35 : La barre d'outils générale de l'explorateur d'archives

Elle présente les fonctionnalités suivantes :

- Version de référence

Cliquer sur ce bouton pour voir la version de référence de la page, c'est à dire la version à laquelle a été comparée la dernière version ce cette page lors du dernier crawl de comparaison.

- Dernière version

Cliquer sur ce bouton pour voir la dernière version archivée.

- Version précédente

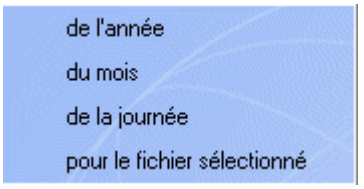
Cliquer sur ce bouton pour voir la version précédente de la page d'archive sélectionnée dans l'explorateur d'archives.

- Version suivante

Cliquer sur ce bouton pour voir la version suivante de la page d'archive sélectionnée dans l'explorateur d'archives.

- Supprimer les archives …

En cliquant sur ce bouton, un sous-menu apparaît :



Les fonctionnalités de ce sous-menu sont identiques à celles proposées par le menu contextuel de l'explorateur d'archives.

- Regroupement

Lorsque l'on clique sur le bouton « Regroupement », un sous-menu apparaît.

Par défaut, l'explorateur d'archives regroupe les archives par année, mois, jour. On peut changer ce regroupement en cochant ou décochant les cases de ce sous-menu :

L'effet est immédiatement visible dans l'explorateur d'archives :

5.1.3 Le browser d'archives

Le browser d'archives permet de visualiser les pages archivées dans la base de données de KB Crawl avec les éventuels surlignements.
Une fois la page affichée, il se comporte comme le navigateur par défaut de l'ordinateur et offre les mêmes fonctionnalités accessibles depuis le menu contextuel de ce navigateur.

Par exemple :
- Afficher la source
- Imprimer
- Exporter vers Excel
- Etc.

Le clic droit sur un lien hypertexte offre les fonctionnalités suivantes :

- Ouvrir dans une nouvelle fenêtre : permet d'explorer une page indiquée par le lien dans le navigateur par défaut de Windows
- Enregistrer la cible sous
- Imprimer la cible
- Etc.

6 Surveillance automatique

6.1 Le mode automatique

Pour effectuer le crawl de comparaison et détecter d'éventuels changements au sein d'une source, KB Crawl propose deux techniques différentes :

- Il est possible de lancer manuellement le crawl de comparaison.

Cette méthode requiert l'intervention de l'utilisateur pour lancer le traitement, ce qui convient pour une comparaison immédiate et ponctuelle.

- Grâce au mode « surveillance automatique », KB Crawl peut également surveiller périodiquement les sources qui ont été programmées à cet effet.

Pour passer en mode « surveillance automatique », il suffit de cliquer sur le bouton « Automatique » de la barre d'outils générale. 

6.2 Paramétrage

Le paramétrage des heures de déclenchement automatique des crawls peut se faire à deux niveaux : celui des sources et celui des dossiers. Ainsi, toutes les sources appartenant à un dossier « héritent » des heures de déclenchement de ce dossier et des éventuels dossiers « parents », s'il y a plusieurs niveaux de dossiers.

Afin de planifier les surveillances automatiques, KB Crawl propose un module de gestion du planning de la surveillance automatique qui permet de gérer facilement les heures de déclenchement à tous les niveaux et de prévisualiser le planning de surveillance résultant de cette gestion.

6.2.1 Accès

Pour accéder à ce module :

- Depuis la barre de menu générale, avec le bouton déroulant « Automatique » puis « Paramètres de surveillance automatique ». Dans ce cas, le module de paramétrage de la surveillance automatique s'ouvre en affichant le planning général contenant toutes les sources à surveiller.

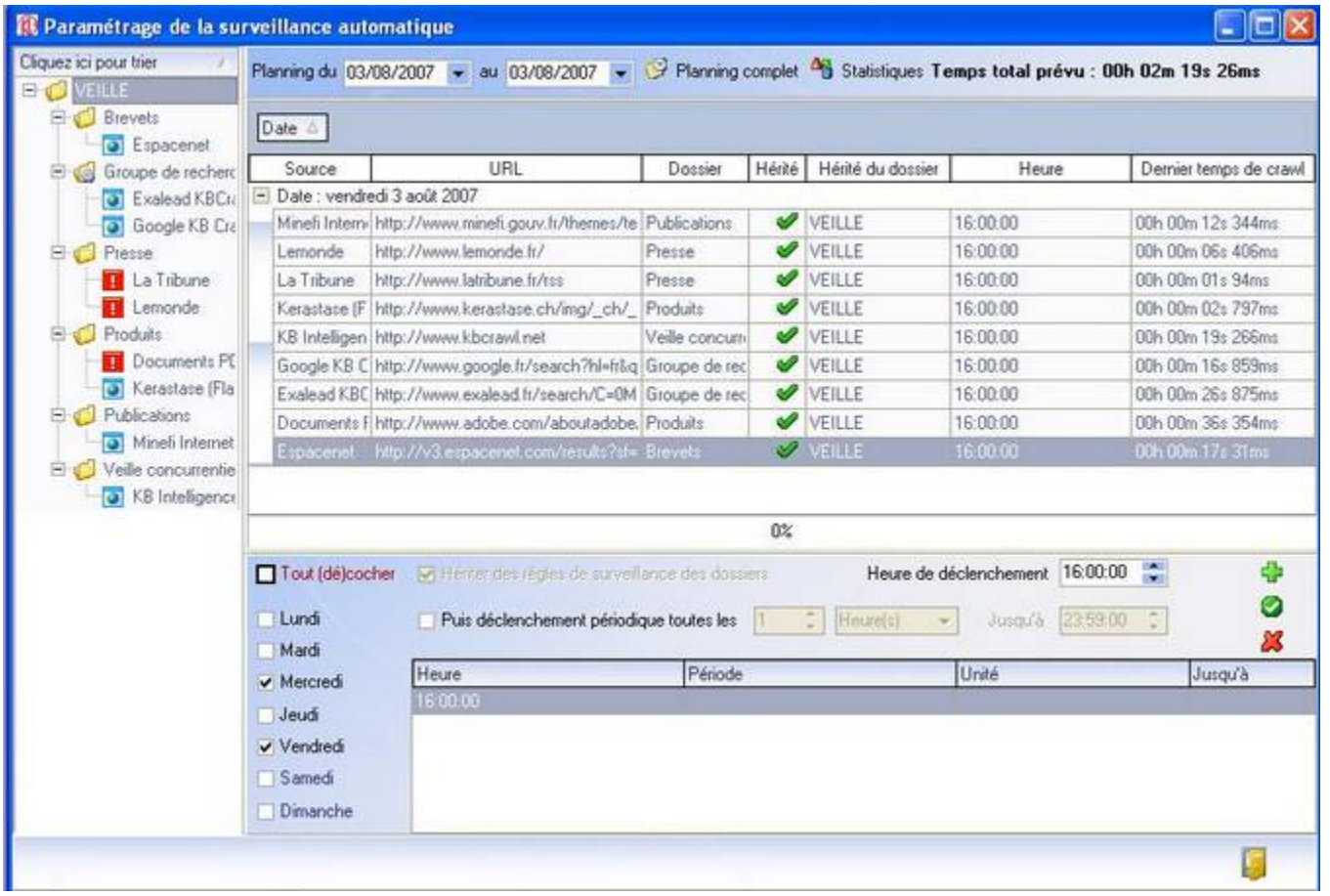


Figure 36 : Affichage du planning complet de surveillance automatique.

- Depuis une source sélectionnée dans le cadre de gauche de la fenêtre principale (clic droit / « Surveillance Automatique » (CTRL + P)). Dans ce cas, le module de paramétrage de la surveillance automatique s'ouvre en affichant le planning de surveillance de la source.

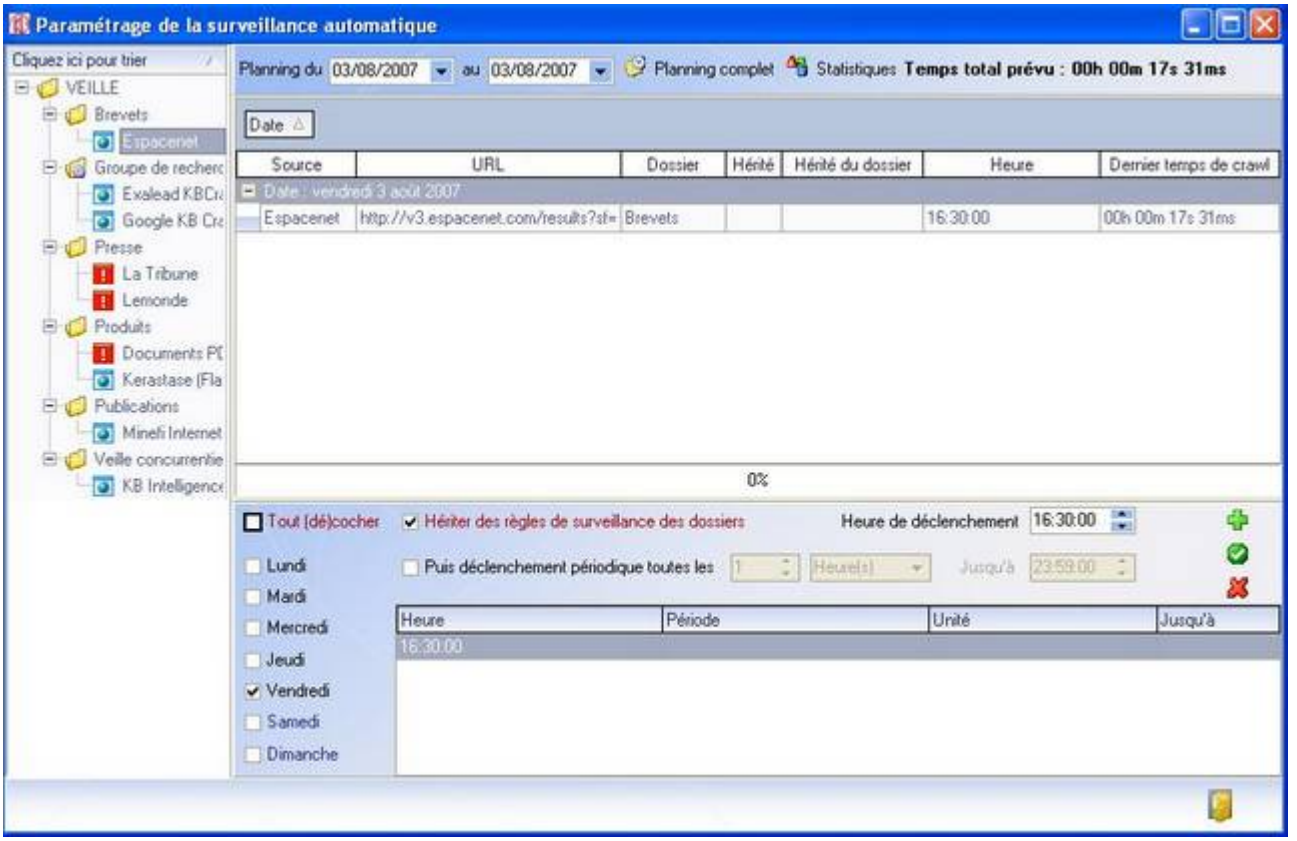


Figure 37 : Affichage du planning de surveillance automatique d'une source.

- De même, un clic droit / Surveillance automatique » (CTRL + P) depuis un dossier ou un sous dossier ouvre le module de paramétrage de la surveillance automatique en affichant le planning de l'ensemble des sources contenues dans ce dossier ou dans un sous dossier de niveau inférieur.

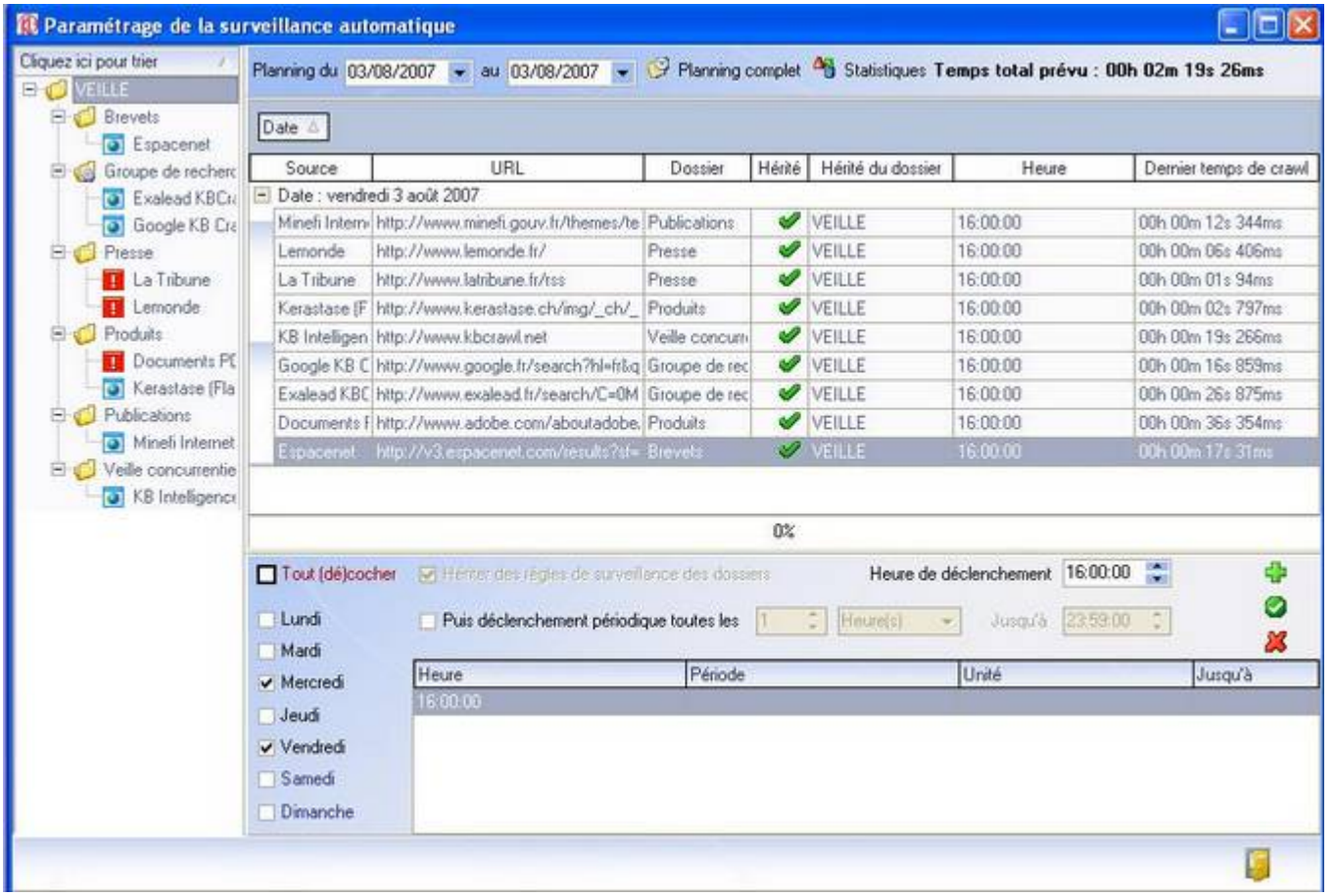


Figure 38 : Affichage du planning de surveillance automatique par dossier.

6.2.2 Ergonomie

Le module de paramétrage présente trois cadres principaux :

- Cadre de gauche

Le cadre de gauche présente la liste des sources et des dossiers sous forme arborescente, similaire en tous points à celui présent dans la fenêtre principale.
On peut ainsi sélectionner la source ou le dossier pour lesquels on souhaite paramétrer le planning de surveillance automatique.

- Cadre du haut

Le cadre du haut présente principalement une grille qui affiche le planning de surveillance automatique d'un élément sélectionné (dossier ou source) ou bien le planning complet.

Cette grille possède 8 colonnes :

- Date
- Source
- URL
- Dossier
- Hérité
- Hérité du dossier
- Heure
- Dernier temps de crawl.

Au sommet de la grille se trouve un panneau de regroupement : on peut glisser/déplacer chacun des entêtes de colonnes pour effectuer un groupement.

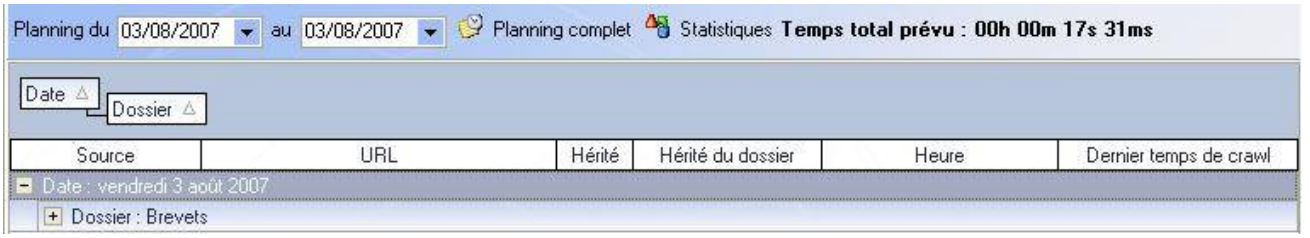


Figure 39 : Planning regroupé, exemple 1.

Il est possible d'effectuer n'importe quel regroupement souhaité.

Par défaut, le planning est présenté avec une rupture par date, uniquement.

Placé au dessus de la grille du planning, un panneau présente plusieurs éléments :

- deux boites de saisie pour spécifier la date de début et la date de fin pour l'affichage du planning (dans la figure ci-contre, on affiche un jour de planning)

- un bouton « Statistiques »

La fonctionnalité « Statistiques » permet d'afficher un graphique qui représente la durée totale des crawls programmés en fonction des heures de la journée.

Selon la durée totale des crawls, cette durée est exprimée en minutes ou en secondes.

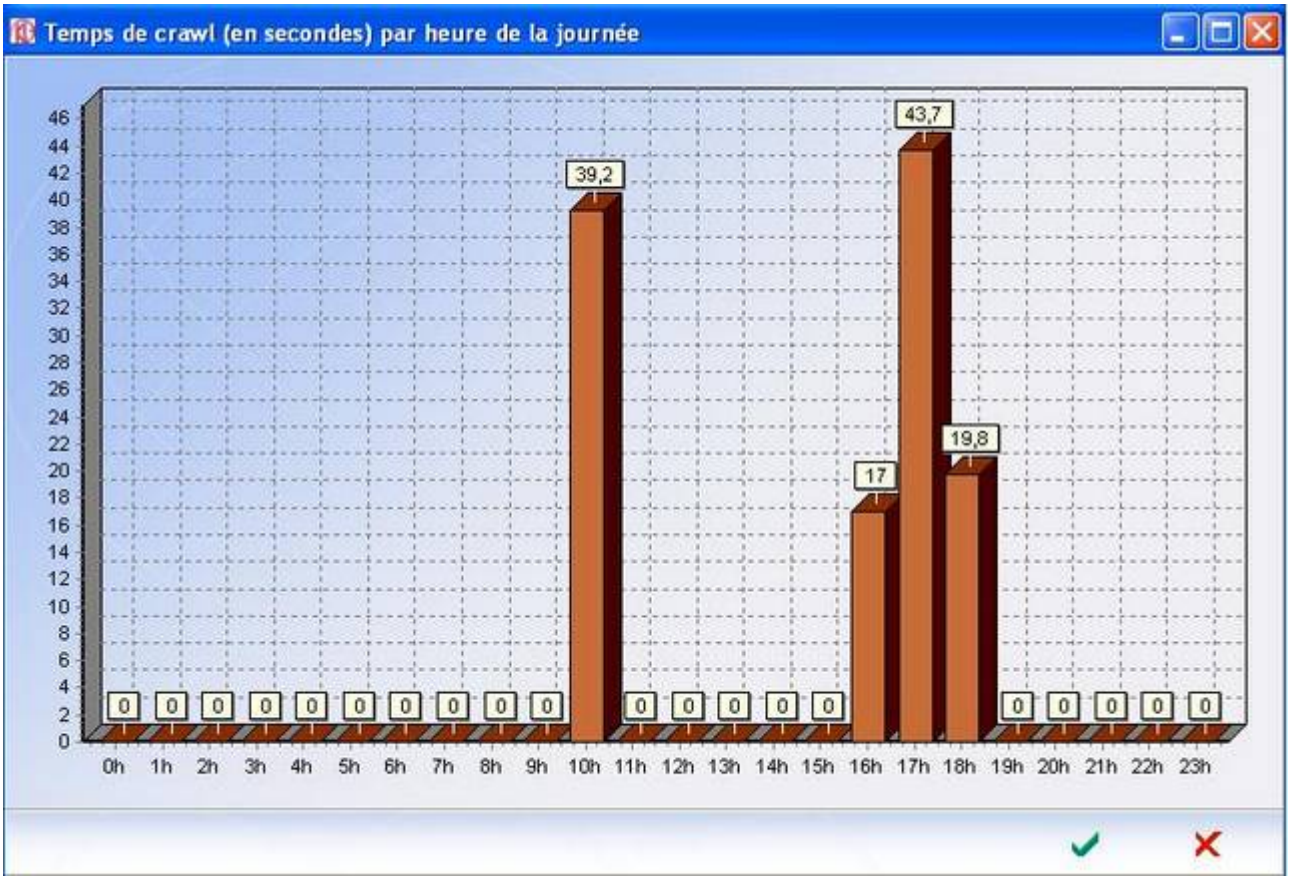


Figure 40 : Durée des crawls programmés en fonction des heures de la journée exprimée en secondes.

Dans ce cas de figure, on voit très clairement que la répartition des crawls est concentrée sur la seconde partie de journée.
Ici, ce n'est pas gênant car la durée totale des crawls pour une heure donnée n'est pas importante.

KB Crawl a la capacité de crawler un très grand nombre de pages et dans le cadre d'un usage « industriel », il est important de répartir correctement les crawls pour ne pas créer de retard dans l'exécution des tâches.

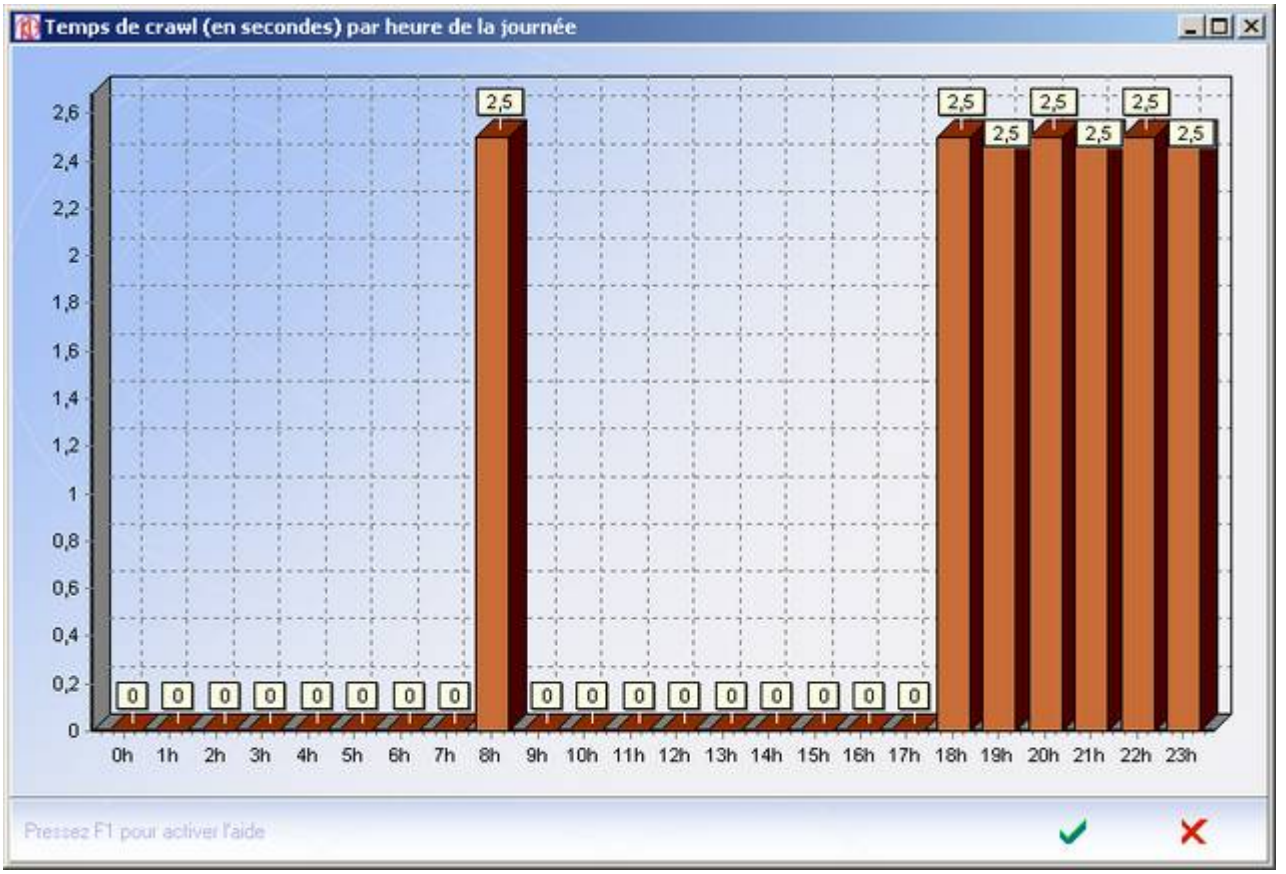


Figure 41 : Durée des crawls programmés en fonction des heures de la journée exprimée en secondes

Dans ce cas précis, seule une source était sélectionnée et son planning affiché. Le graphique représente la répartition des crawls qui ont été programmés pour une seule source pour se déclencher une fois à 8h00 puis toutes les heures jusqu'à 23h59.
La durée des crawls est exprimée en secondes.

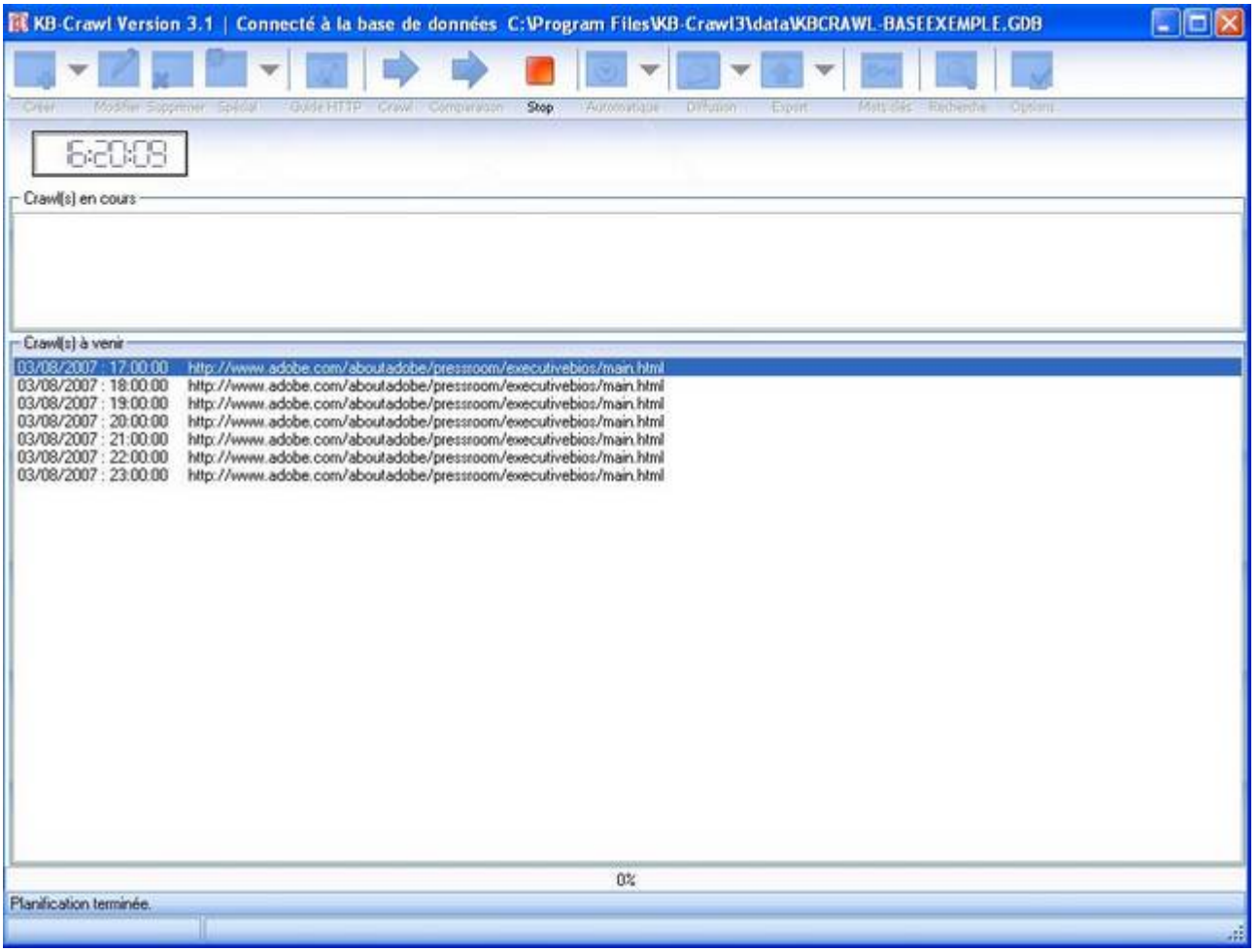
Le dernier élément, à droite du bouton « statistiques » est le « temps total prévu ».
La valeur affichée est le temps total prévisionnel des crawls affichés dans la grille du planning.
Le calcul se base sur les dernières durées de crawl constatées pour chaque source.

Si cette durée dépasse 24h, la valeur s'affiche en rouge à titre d'avertissement. Dans ce cas en effet les crawls « à venir » risquent de s'empiler dans la file d'attente et de ne pas être traités à temps.

- Cadre du bas

Le cadre du bas sert à ajouter, modifier ou supprimer des heures de déclenchement.

Une fois le mode « surveillance automatique » enclenché, la liste des sources et de leurs heures de déclenchement programmées pour le jour en cours s'empilent selon l'ordre chronologique croissant, puis alphabétique dans l'espace « Crawl(s) à venir ».



Cette pile de sources à crawler ainsi constituée est inspectée à chaque seconde par l'horloge pour faire passer les éléments de cette pile dont l'heure de déclenchement est d'actualité dans l'espace « En cours ».

L'espace « En cours » est une file d'attente dans laquelle chaque source « attend son tour » pour un crawl de comparaison.
Chaque jour, à minuit, l'espace « A venir » est actualisé en fonction des sources pour lesquelles une programmation de surveillance automatique est prévue pour le jour qui commence.

6.2.3 Fonctionnement

6.2.3.1 Affichage du planning

Le planning affiché ne tient pas compte de l'heure en cours au moment de la consultation, il sert à prévisualiser le planning de la surveillance automatique pour une journée donnée.
A chaque fois qu'un paramètre est modifié sur un des cadres (plage de dates, heures et jours de déclenchement, sélection d'une source ou d'un dossier), le planning est automatiquement recalculé et rafraîchi.

6.2.3.2 Héritage

Par défaut, une source hérite des heures de déclenchement de la hiérarchie de dossiers dans laquelle elle est contenue :

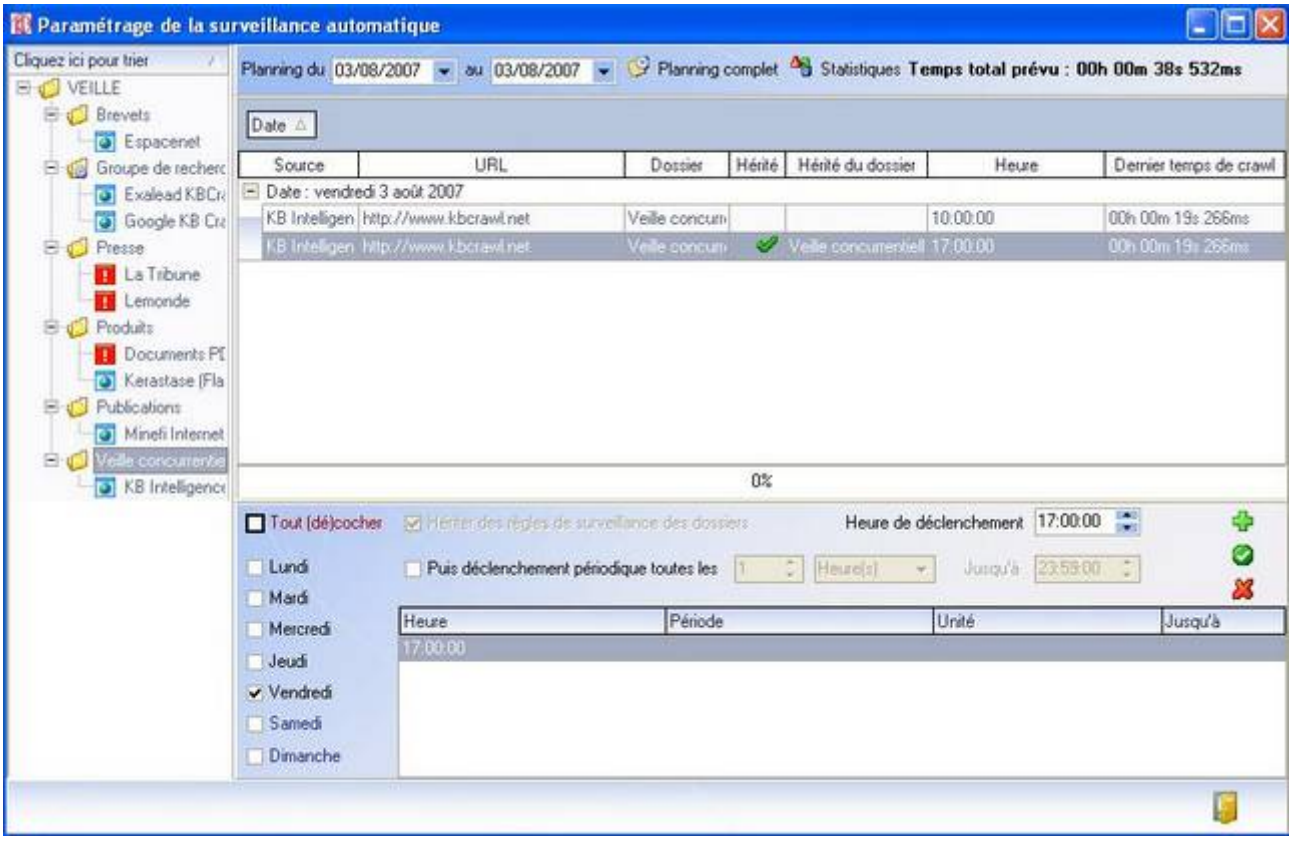


Figure 42 : Héritage des heures de déclenchement.

Dans le cas illustré ici, une seule heure de déclenchement a été programmée pour la source sélectionnée. Cependant, le planning affiche deux heures de déclenchement. Une des heures de déclenchement est héritée du dossier qui contient la source (ce qui est clairement notifié dans la colonne « hérité »).
Si l'on décoche l'option « Hériter des règles de surveillance des dossiers », les autres heures que 10h00 disparaissent du planning :

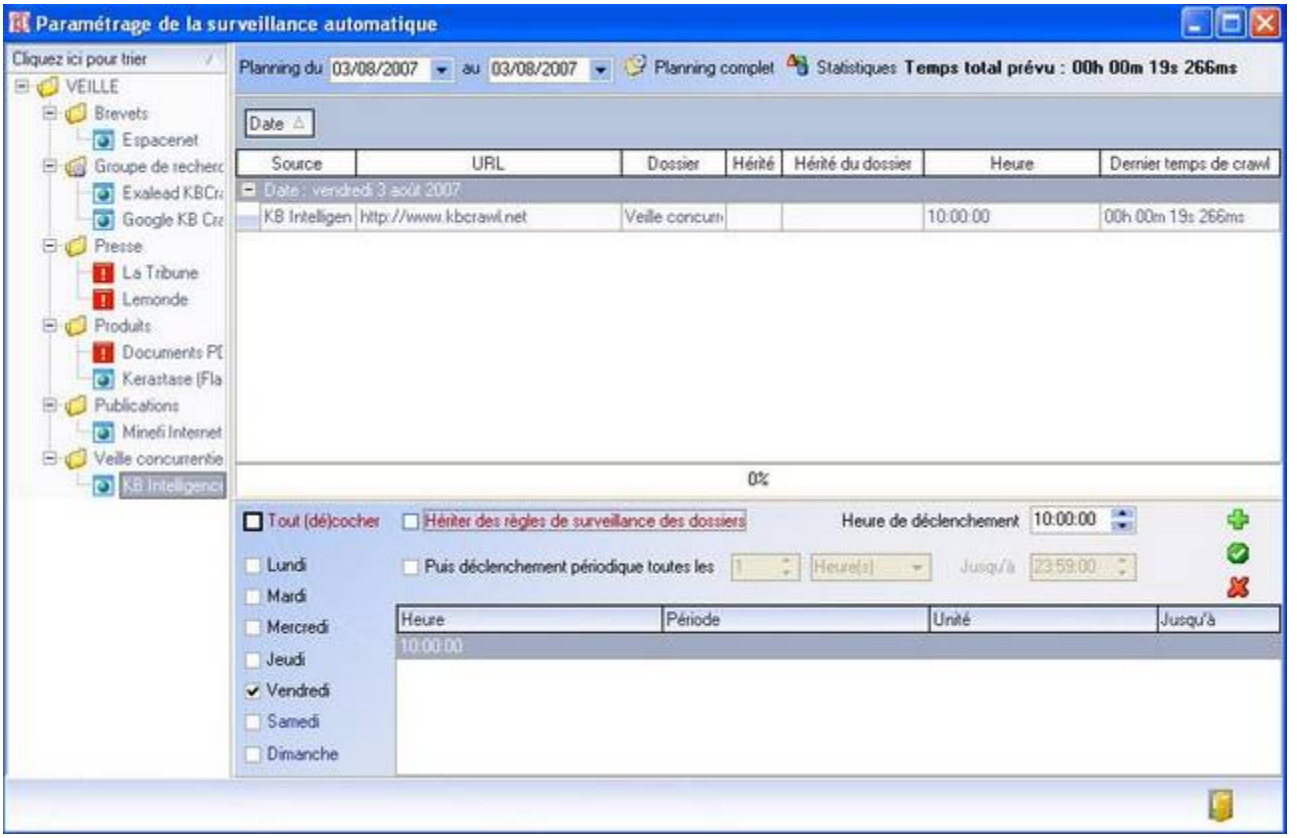


Figure 43 : Non héritage des heures de déclenchement.

L'horaire 17h00 présent dans le planning d'exemple précédent a été programmé pour le dossier « Veille concurrentielle»
Il est possible de le visualiser en cliquant sur un des dossiers dans le cadre de gauche.

Autre exemple :

On est vendredi. Si on décoche le Vendredi dans le cadre du bas, en étant placé sur la source :

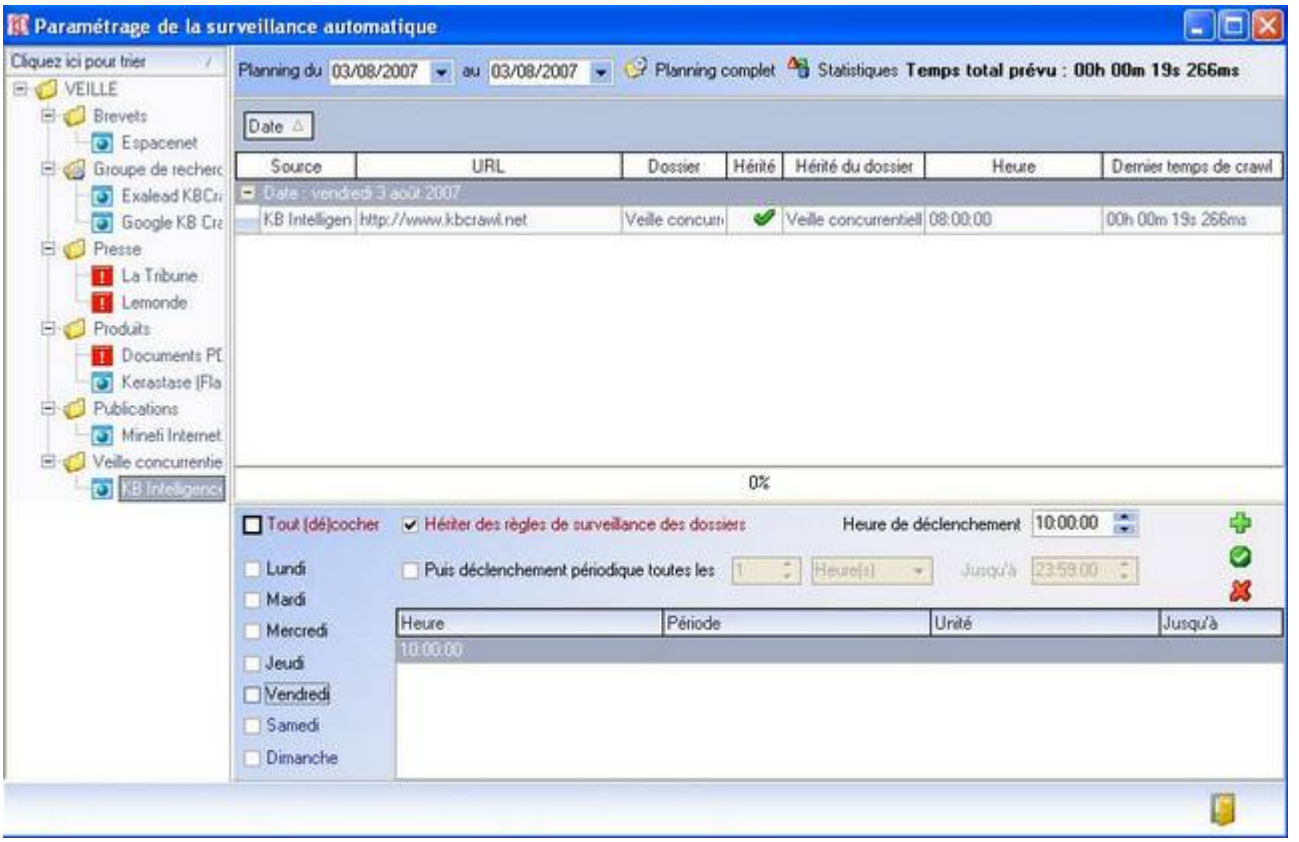


Figure 44 : Héritage de surveillance automatique, autre exemple

L'horaire 10h00 disparaît alors que les autres horaires persistent : les jours actifs du cadre du bas sont contextuels à l'entité sélectionnée. Ici, ils ne concernent que les heures de déclenchement de la source.

Pour les dossiers pères, tous les jours sont cochés, ce qui explique que les autres horaires liés aux dossiers persistent dans le planning.

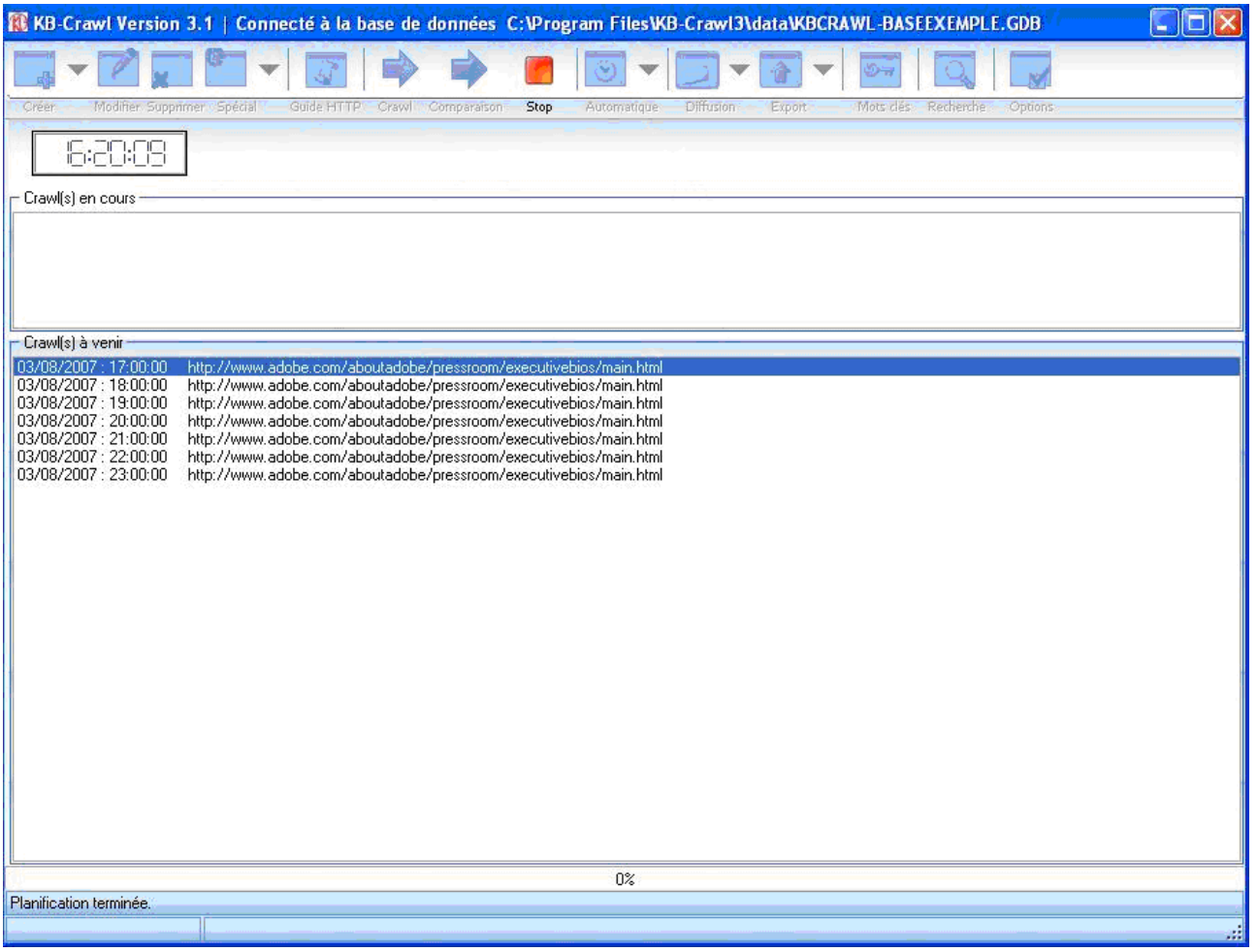


Figure 45 : L'onglet "automatique" avec surveillance automatique enclenchée.

Lorsque le mode « surveillance automatique » est enclenché, aucune fonctionnalité de KB Crawl n'est accessible, il faut désenclencher ce mode pour y avoir de nouveau accès.

Pour désenclencher le mode automatique, il suffit de cliquer sur le bouton d'arrêt : 
Attention : Cette action entraîne l'annulation du crawl éventuellement en cours.

6.3 Le démarrage en mode automatique

KB Crawl peut être démarré en mode automatique. Pour ce faire, il faut exécuter le fichier « crawler.exe » avec le paramètre « AUTO ». Le fichier « crawler.exe » se trouve dans le répertoire d'installation de l'application.
Il suffit pour cela de créer un raccourci qui pointe vers « crawler.exe » avec le paramètre « AUTO ».

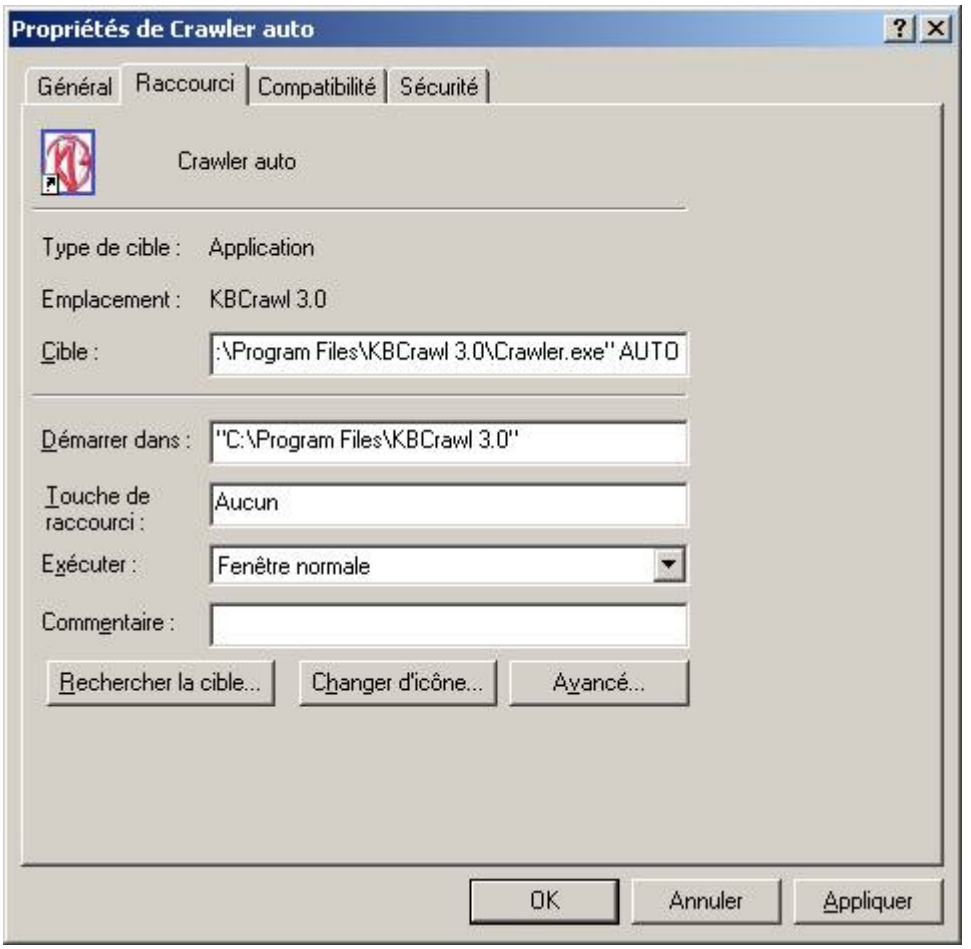


Figure 46 : Création d'un raccourci pour lancer KB Crawl en mode automatique

Lorsque KB Crawl est lancé en mode automatique, il attend 15 secondes environ avant de s'instancier. Il est donc normal de ne pas voir l'application démarrer instantanément.

7 La liste de diffusion

Lorsque KB Crawl déclenche une alerte sur une source, il peut envoyer un message par e-mail à un ou plusieurs contacts dont on connaît l'adresse électronique.
Ces contacts peuvent être ajoutés à une liste (= carnet d'adresses) et être regroupés dans des groupes de contacts.

7.1 Liste des contacts

Pour accéder à la liste des contacts, cliquer sur le bouton « Diffusion » de la barre d'outils générale, puis dans le menu déroulant, choisir « Contacts ».

La liste des contacts apparaît :

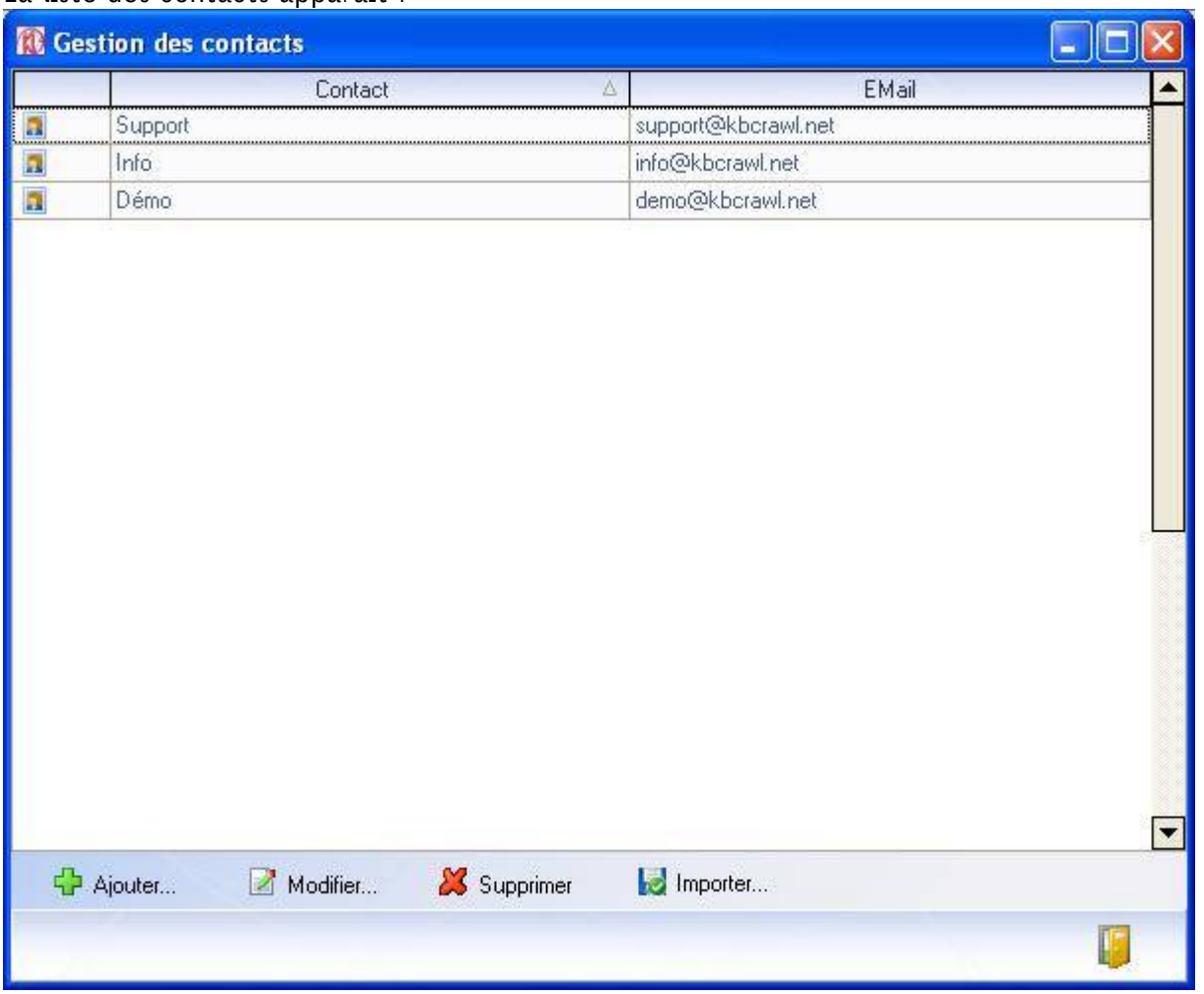


Figure 47 : Liste des contacts.

Ajouter un contact

Cliquer sur le bouton « Ajouter», puis choisir le type de contact : Individu ou Groupe.

Renseigner les champs Noms, Prénoms et E-mail du contact puis valider en cliquant sur ce bouton .

Modifier un contact
Cliquer sur le bouton « Modifier »

Figure 48 : Détail d'un contact

Modifier ensuite les champs Nom, Prénom et E-mail du contact puis valider en cliquant sur ce bouton .

Supprimer un contact

Cliquer sur le bouton supprimer 

Puis valider en cliquant sur « Oui ».

Importer des contacts

Il est possible d'importer une liste de contacts e-mails à partir d'un fichier Microsoft Excel ou CSV.
Chaque ligne doit contenir 3 colonnes : la première doit contenir l'adresse e-mail, la deuxième, le nom du contact et la troisième le prénom du contact.
Les noms et prénoms peuvent être omis. L'import s'arrête lorsqu'une ligne ne contient pas d'adresse e-mail.

Note : Dans le cas du fichier Excel, l'import sait gérer les fichiers XLS exportés à partir de Microsoft Outlook.

7.2 Groupes de contacts

Ajouter un groupe

Dans la liste des contacts, cliquer sur le bouton « Ajouter » puis choisir « Groupe » dans la boîte de dialogue suivante.

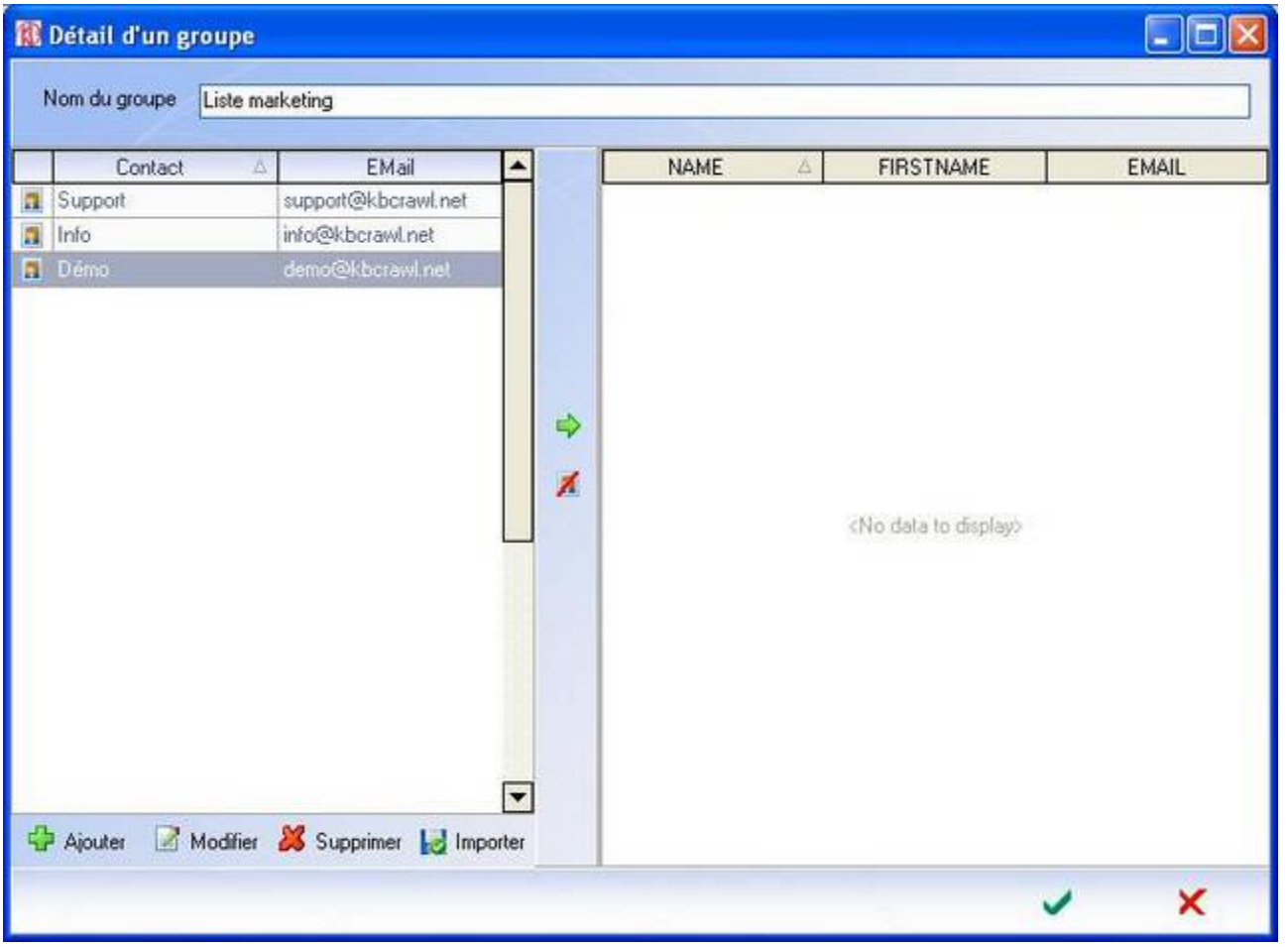


Figure 49 : Détail d'un groupe de contacts

L'écran de gestion d'un groupe comporte deux fenêtres séparées par une colonne présentant deux boutons.

- La fenêtre de gauche contient la liste des contacts disponibles.
- La fenêtre de droite contient la liste des membres du groupe.
- Le premier bouton de la colonne centrale permet d'ajouter le contact sélectionné dans la fenêtre de gauche aux membres du groupe.
- Le second bouton permet d'ôter le membre sélectionné dans la fenêtre de droite du groupe.

7.3 Gestion des abonnements

Après avoir créé les contacts et les groupes de contacts, il suffit, pour que ceux-ci reçoivent des alertes par e-mail de les abonner à des dossiers.
Ainsi, lorsqu'un contact est abonné à un dossier et qu'une source appartenant à ce dossier fait l'objet d'une alerte, il reçoit automatiquement un message par courrier électronique.
Pour gérer ces abonnements, cliquer sur « Abonnements » dans le menu déroulant de « Liste de diffusion ».

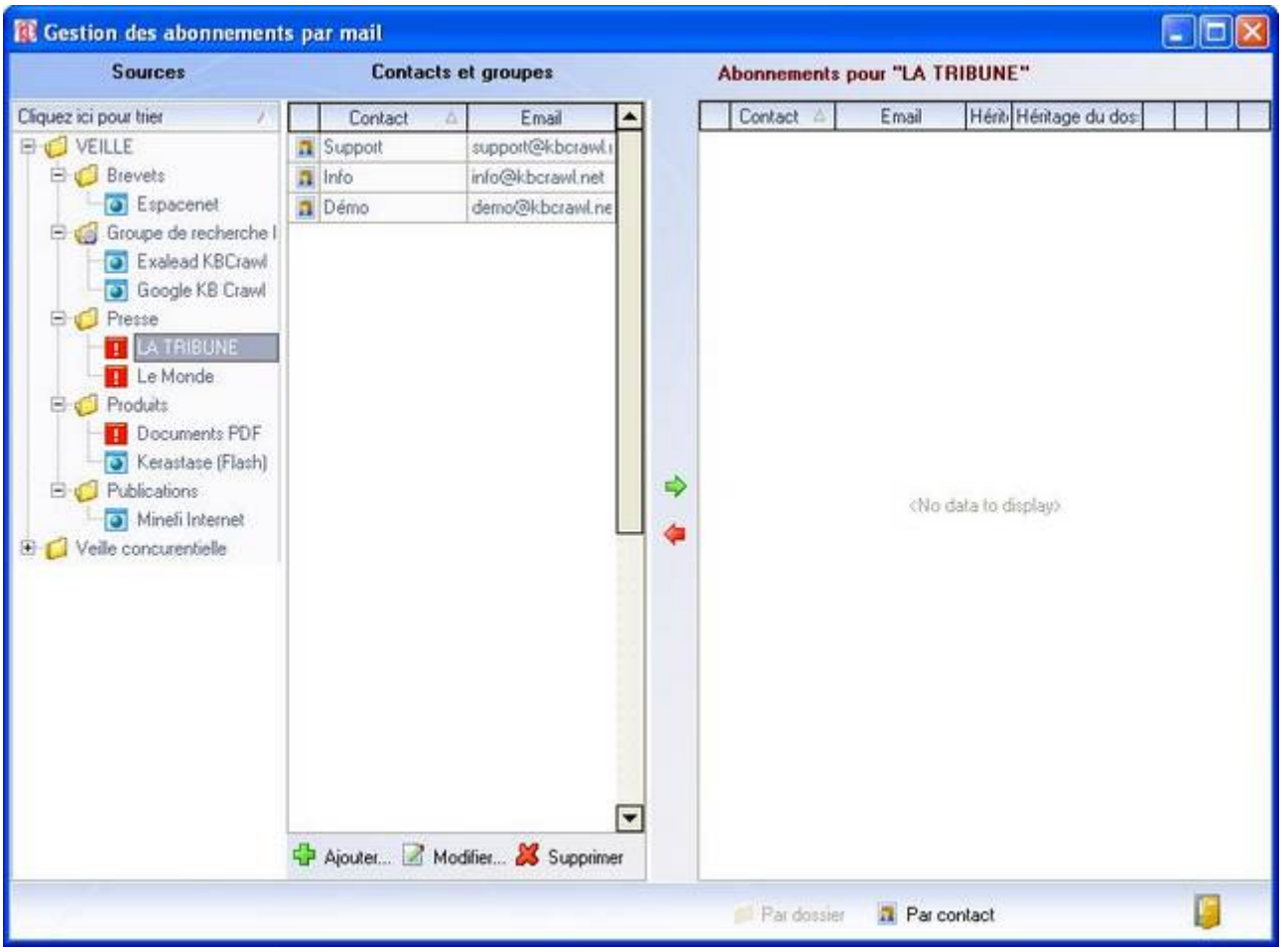


Figure 50 : La gestion des abonnements

L'écran de gestion des abonnements comporte trois cadres :

- Cadre de gauche

Il contient l'arborescence des sources et des dossiers.

- Cadre du milieu

Il contient la liste des contacts, avec les mêmes fonctionnalités que dans le menu « gestion des contacts ».

- Cadre de droite

Il contient la liste des contacts abonnés à la source ou le dossier sélectionné(e) dans le cadre de gauche.

Le bouton de la colonne centrale permet d'abonner un contact ou un groupe de contacts à la source ou au dossier sélectionné(e).

Le second bouton permet de supprimer l'abonnement d'un contact ou d'un groupe de contacts.

7.3.1 Ajout d'un abonné

Cliquer sur le premier bouton de la colonne centrale 

Une fenêtre apparaît :

Figure 51 : Détail d'un abonnement.

L'alerte pour une source donnée peut se produire pour différentes raisons (changements dans le contenu, apparition de mots-clés, page disparue, etc.) mais chaque abonné ne souhaite pas forcément être alerté pour n'importe laquelle de ces raisons. Un contact peut souhaiter n'être alerté que pour un motif bien précis.

Pour cela, KB Crawl permet pour un même contact ou groupe de contacts de personnaliser ou filtrer l'alerte qu'il reçoit.

Dans cet exemple, le nouvel abonné ne recevra un message par e-mail que si des mots-clés apparaissent sur une des pages de la source.

7.4 Gestion des e-mails en attente

Les messages d'alerte peuvent être envoyés automatiquement après chaque crawl ou comparaison.
Si cette option n'est pas retenue, les messages sont stockés dans une boîte d'envoi et peuvent être envoyés manuellement à tout moment.
Pour visualiser le contenu de cette boîte d'envoi, cliquer sur « Gestion des mails en attente » dans le menu déroulant de « diffusion ».

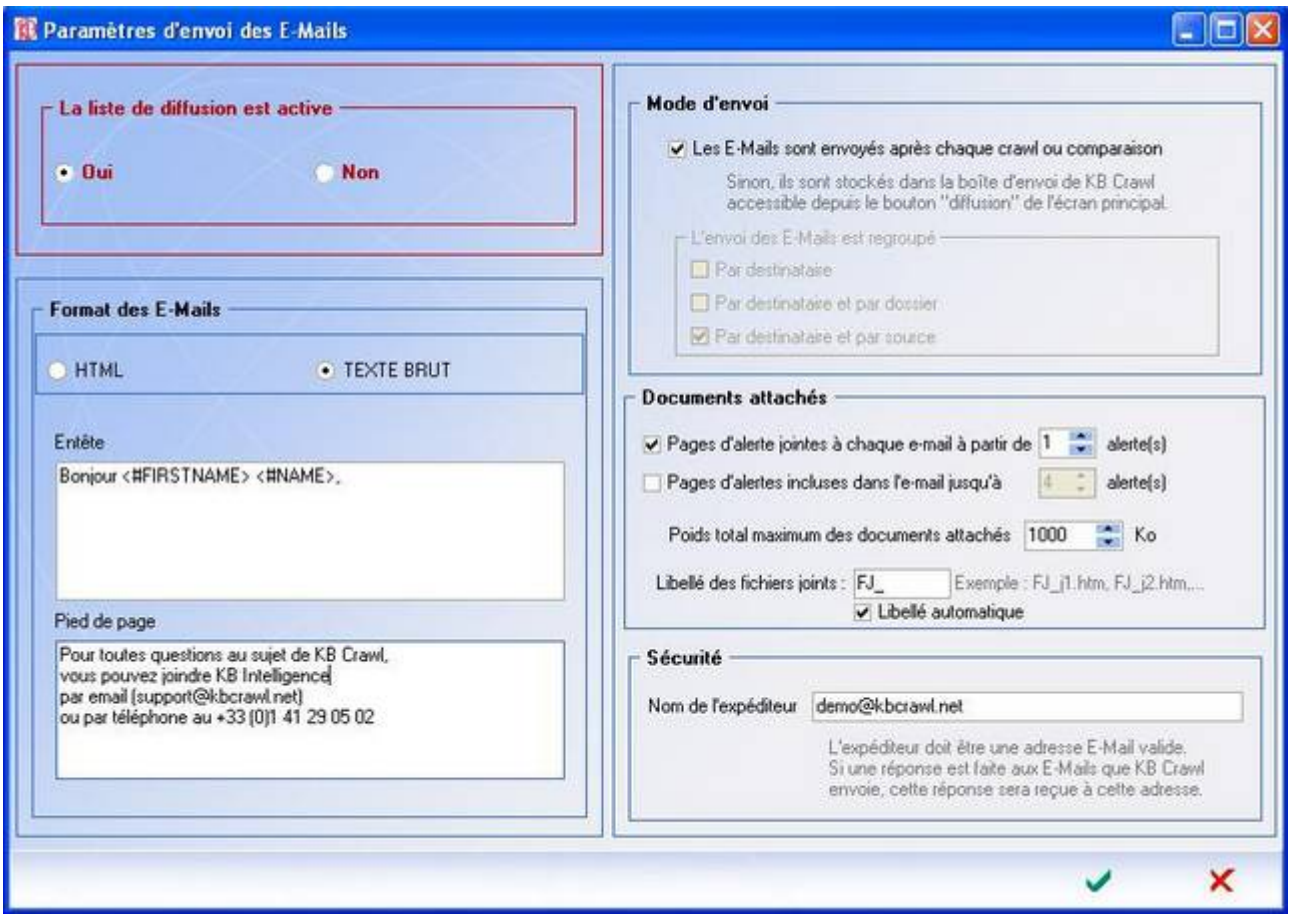
Figure 52 : Liste des messages à envoyer.

La liste des messages à envoyer montre pour chaque contact la liste des messages qui lui sont destinés.
On peut aussi regrouper ces messages par dossier, source, nom et prénom.

Ce module permet également de vider la boîte d'envoi de tous les messages ou simplement de tous les messages du contact sélectionné.
Pour cela, faire un clic droit depuis le contact sélectionné ou cliquer sur le bouton supprimer de la barre d'outils située en haut de l'écran, puis cliquer sur l'élément de menu adéquat.

7.5 Paramètres d'envoi

Il est possible de modifier les paramètres d'envoi des e-mails d'alerte envoyés par KB Crawl et par conséquent, de personnaliser la diffusion des e-mails en modifiant les paramètres d'envoi.



Il est possible d'agir sur tous les paramètres suivants :

- La liste de diffusion est active

Cette option permet de choisir ou non de prendre en compte les abonnements créés dans le menu « Gestion des abonnements ». Si la case « Non » est cochée, cela signifie que KB Crawl ne prendra en compte aucun des abonnements créés et ne diffusera aucun e-mail d'alerte.

- Format des E-Mails

Il est également possible de choisir le format d'envoi des e-mails générés par KB Crawl. Deux formats sont proposés : HTML et Texte Brut.

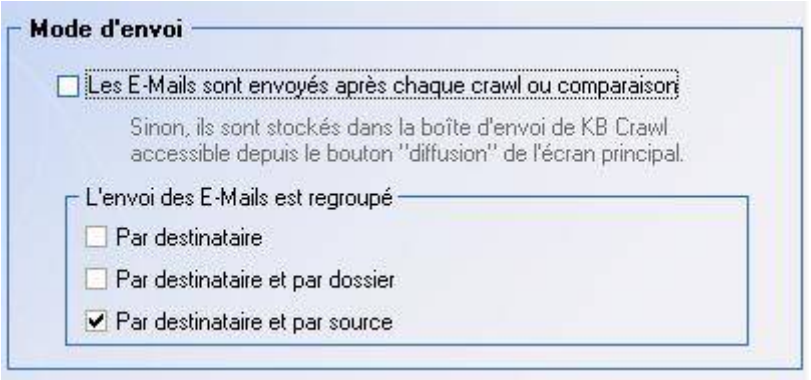
- Dans le cas d'un envoi au format HTML, KB Crawl s'appuiera sur un fichier HTML comme « modèle » pour le corps du mail ; ce fichier est stocké dans le dossier d'installation par défaut de KB Crawl et se nomme alerte.htm

- Dans le cas d'un envoi au format texte brut, un entête et un pied de page sont proposés par KB Crawl ; ils apparaîtront respectivement en haut et en bas de chaque e-mail d'alerte envoyé par KB Crawl. Ces informations sont modifiables directement dans cette fenêtre.

- Mode d'envoi

Par défaut, les e-mails d'alerte sont envoyés par KB Crawl directement après chaque crawl ou comparaison, cependant, il est possible de choisir de ne pas les envoyer automatiquement. En décochant la case « les E-Mails sont envoyés après chaque

crawl ou comparaison », les e-mails d'alerte seront stockés dans la boîte d'envoi de KB Crawl (Cf 8.4). De plus, si cette case est décochée, il est possible de sélectionner le type de regroupement que l'on souhaite appliquer aux e-mails d'alerte qui seront envoyés par KB Crawl.



Il existe trois types de regroupement possibles :

- Par destinataire : chaque destinataire recevra un seul et unique e-mail d'alerte par session de crawl (indépendamment du nombre de sources en alerte).

- Par destinataire et par dossier : chaque destinataire recevra un e-mail par dossier contenant au moins une source en alerte pour laquelle il est abonné. Cela signifie qu'une personne abonnée à 4 sources appartenant à 3 dossiers différents ne recevra que 3 e-mails d'alerte.

- Par destinataire et par source : chaque destinataire recevra autant d'e-mails qu'il existe de sources en alerte pour lesquelles il est abonné. Cela signifie que si l'on reprend l'exemple précédent, le destinataire recevra 4 e-mails d'alerte.

- Documents attachés

KB Crawl offre la possibilité d'envoyer, en pièce jointe des e-mails d'alerte, une copie des pages en alerte, ainsi que les inclure dans le corps du mail.

- Pour joindre les pages d'alerte aux e-mails, il suffit de cocher la case « Pages d'alerte jointes à chaque e-mail à partir de » et de préciser une valeur minimale. Si

la valeur est 1 (valeur par défaut), cela signifie que les pages d'alertes seront toujours envoyées en pièces jointes des e-mails.

- Il est aussi possible d'inclure les pages d'alerte dans le corps des e-mails ; pour cela, il suffit de cocher la case « Pages d'alerte incluses dans l'e-mail jusqu'à » et de préciser une valeur maximale. Au-delà de cette valeur, les fichiers ne seront plus dans le corps de l'e-mail.
Il est possible de combiner des fichiers en pièce jointe ainsi que des alertes dans les corps de l'e-mail.
ATTENTION : l'inclusion des pages dans le corps de l'e-mail n'est possible que si le format choisi pour la diffusion des e-mails est HTML.

- Le poids total maximum des documents attachés correspond à la valeur maximale en Ko qu'il est possible d'atteindre pour un e-mail d'alerte. Au-delà de ce poids, les fichiers d'alerte ne seront plus attachés à l'e-mail (mais celui-ci sera tout de même envoyé).

- Il est possible de choisir le libellé des fichiers joints aux e-mails d'alerte en renseignant le champ « Libellé des fichiers joints ». Dans le cas où l'on souhaiterait un libellé automatique différent à chaque session d'envoi, il suffit de cocher la case « Libellé automatique ».

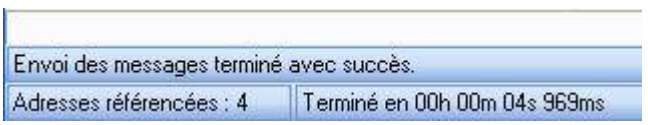
- Sécurité

Lorsque KB Crawl envoie des e-mails d'alerte, l'information se trouvant dans le champ « Nom de l'expéditeur » apparaîtra dans le champ « De » ou « From » de l'e-mail envoyé.

7.6 Envoi des messages

L'envoi des messages se fait soit automatiquement après chaque crawl, soit manuellement en cliquant sur « Envoyer les messages » dans le menu déroulant de « Diffusion/Gestion des E-Mails en attente ».
La barre d'état indique alors que l'envoi des messages est en cours et la barre de progression renseigne sur la progression de l'envoi des messages.

A la fin de l'envoi, elle renseigne sur le résultat de l'envoi :



Il se peut que l'envoi échoue, pour un problème lié au serveur de messagerie ou pour un problème de paramétrage SMTP.

Dans ce cas, un message d'information apparait dans la barre d'état :

Figure 53 : Erreur lors d'envoi de messages.

8 Fonctions d'export

Tous les documents extraits du Web par KB Crawl sont stockés dans sa base de données et consultables depuis l'interface de KB Crawl, notamment avec l'explorateur d'archives.

Tous ces documents stockés au format HTML sont récupérables à n'importe quel moment grâce à la fonction d'export. On appelle export le fait d'extraire un lot de documents HTML de la base de données pour les placer dans un répertoire du disque dur.

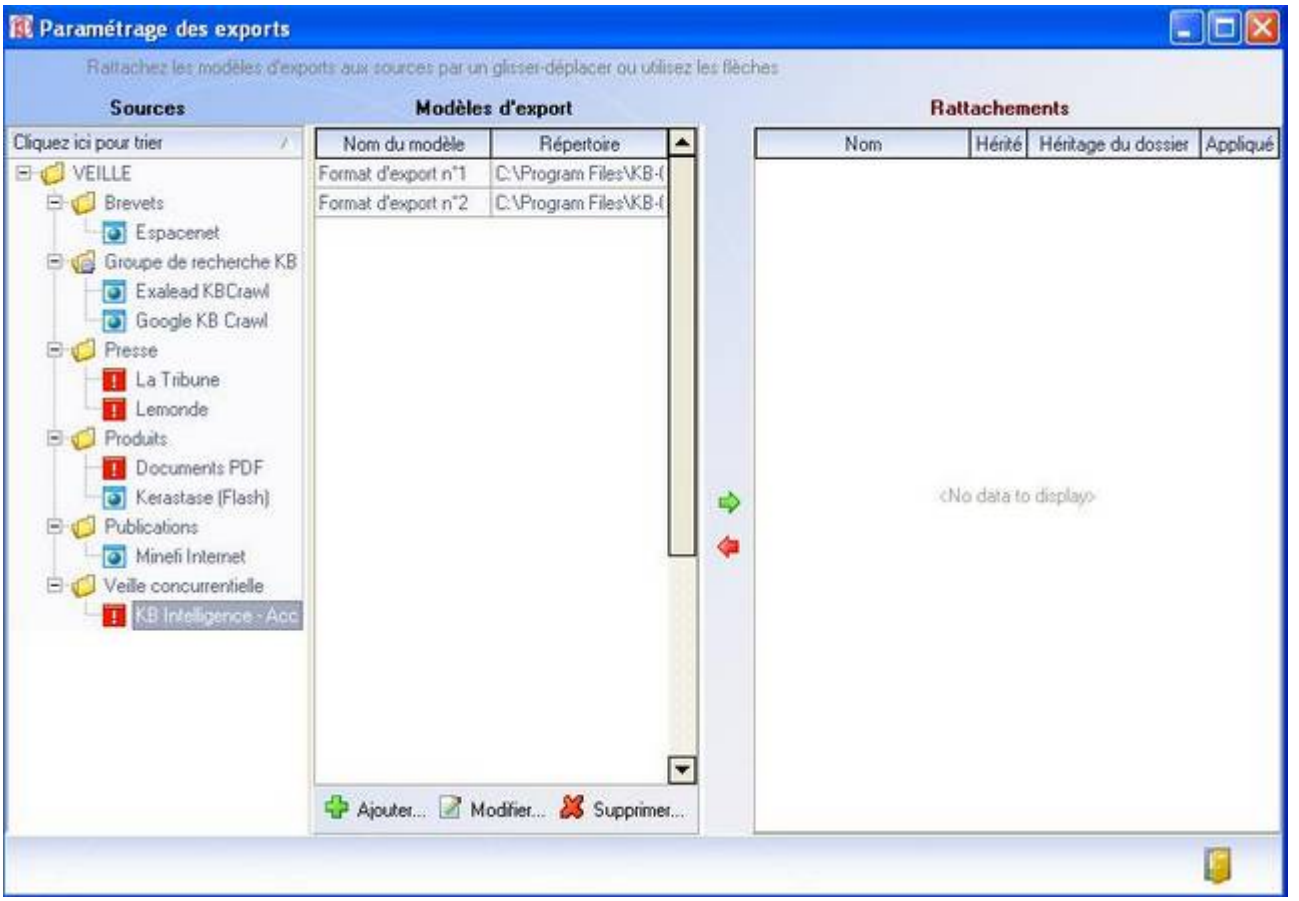
Les documents sont récupérés tels quels, rangés au même niveau dans un dossier et accompagnés d'un fichier d'index au format HTML qui répertorie tous ces documents et permet d'y accéder directement.

Ce fichier contient plusieurs informations pour chaque document répertorié :

- le nom du document, formé de différents éléments qui sont paramétrables,
- la version du document (version de référence, version intermédiaire, dernière version),
- les motifs résumés de l'alerte, s'il y en a une, on retrouve les informations présentes dans l'onglet « liste ».

Tous ces éléments paramétrables constituent le format d'export. Chaque source ou dossier peut être rattaché à un format d'export.

Pour effectuer ce paramétrage, il suffit de cliquer sur le bouton export  depuis la barre d'outils générale.



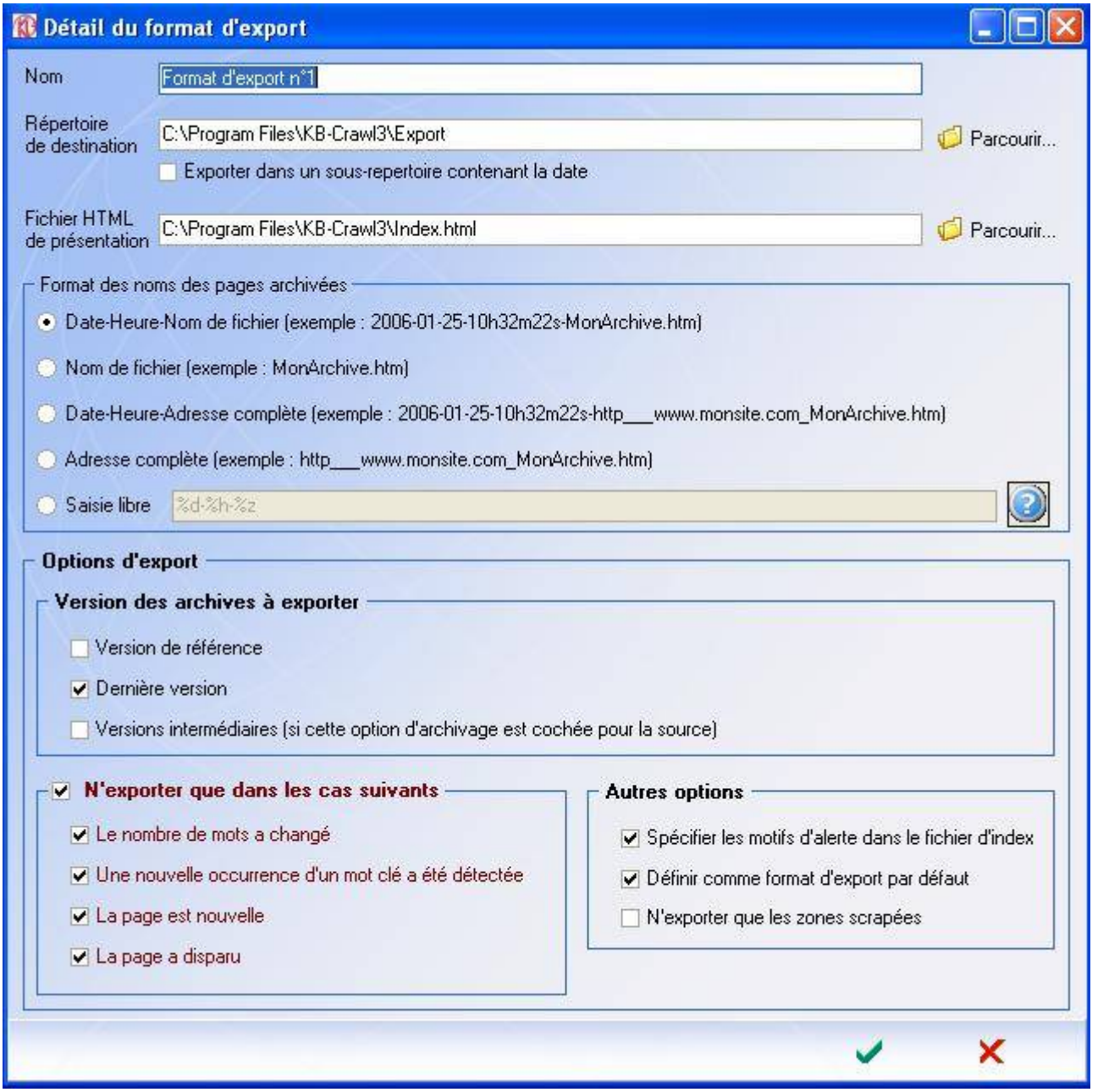
L'écran de paramétrage des exports se divise en trois parties :

8.1 Le cadre de gauche

Le cadre de gauche présente la liste des sources et des dossiers sous forme arborescente, similaire en tous points à celui présent dans la fenêtre principale.
Dans un premier temps, sélectionner la source ou le dossier auquel le modèle d'export va être rattaché.

8.2 Le cadre central

Il dresse la liste des modèles d'exports qui sont déjà paramétrés dans KB Crawl.
Il est possible d'en ajouter. Pour cela, cliquer sur le bouton « ajouter » situé dans la barre de boutons au bas du cadre.



Dans l'ordre, les informations suivantes sont à saisir :

- Nom

  C'est le titre du format d'export, qui permet de le désigner lorsque l'on veut l'utiliser pour un export.

- Répertoire de destination

Tous les fichiers HTML exportés seront placés dans un sous répertoire du répertoire de destination. Ce sous répertoire portera le nom du catalogue ou bien un nom composé ainsi : RechercheDate-Heure si l'export est fait à partir d'une recherche. Par défaut, le répertoire de destination est celui où se trouve l'exécutable de KB Crawl. Mais on peut définir le répertoire que l'on souhaite.

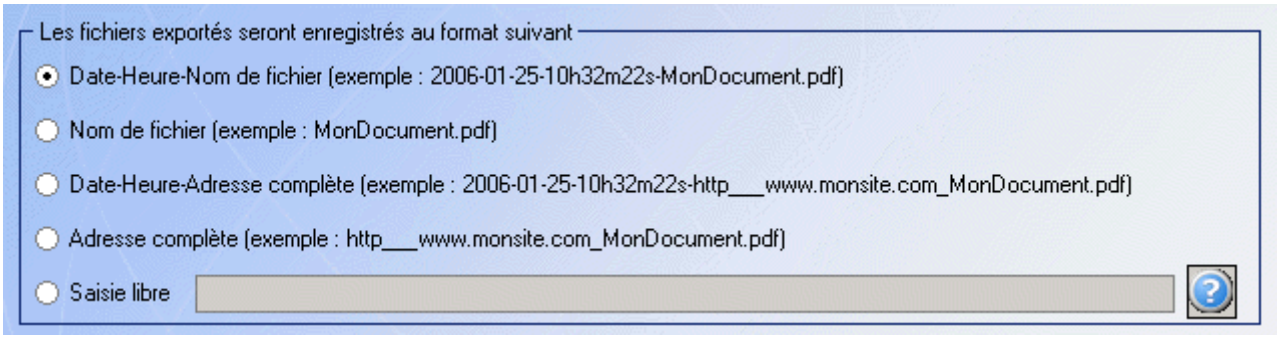
- Fichier HTML de présentation

Cette information concerne le fichier d'index créé au même niveau que les documents exportés.

Un gabarit pour ce fichier est livré et installé avec KB Crawl. Le fonctionnement est le même que pour le fichier HTML qui permet de constituer les e-mails.

Il permet de personnaliser la page d'index autour des parties non modifiables : le titre et la liste des liens, dynamiquement constitués.

- Formats des noms de fichiers



Ensuite, il faut spécifier le format du nom des fichiers enregistrés. Des formats pré établis sont proposés (4 premiers boutons radios) car ils sont pratiques et souvent utilisés, mais il est également possible de composer son propre masque de nom de fichier à partir de variables comme %d par exemple qui prendra la valeur de la date au moment de l'enregistrement du fichier.

Un bouton d'aide permet de voir la liste des variables que l'on peut utiliser :

Exemple de masque : %x_%t _%d_%h pourra donner le nom de fichier suivant :

www.kbcrawl.com_presentation.pdf_15_01_2006_15_05_55

Notons que les caractères « / » et « : » ont été remplacés automatiquement par des « _ » parce qu'ils sont interdits dans les noms de fichiers Windows.

Plusieurs formats pré établis sont proposés dans la liste déroulante.
Celui proposé par défaut est le suivant : %x (%d %h)

Ce qui donne par exemple :

www.kbcrawl.com (28_12_2004 19_11_26)_2.html

Les caractères suivants sont interdits dans les noms de fichiers Windows et seront automatiquement remplacés par des « _ »
\ / : * ? « < > |
D'autre part, les noms de fichiers dépassant 255 caractères seront automatiquement tronqués.

- Version des documents à exporter

A chaque export, KB Crawl exporte les documents qui sont archivés dans la base de données et liés à la source.
En règle générale, on ne souhaite exporter que la dernière version pour chaque document, mais il est possible d'exporter la version de référence de chaque document (à des fins de comparaison par exemple) ou même toutes les versions de chaque document.

- N'exporter que dans les cas suivants

Afin de filtrer d'avantage les documents qui sont exportés à la fin d'un crawl de comparaison, il est possible de jouer sur les critères d'alertes déclencheurs ou non de l'export :

- Le nombre de mots a changé,
- Une nouvelle occurrence d'un mot-clé a été détectée,
- La page est nouvelle,
- La page a disparu.

- Autres options

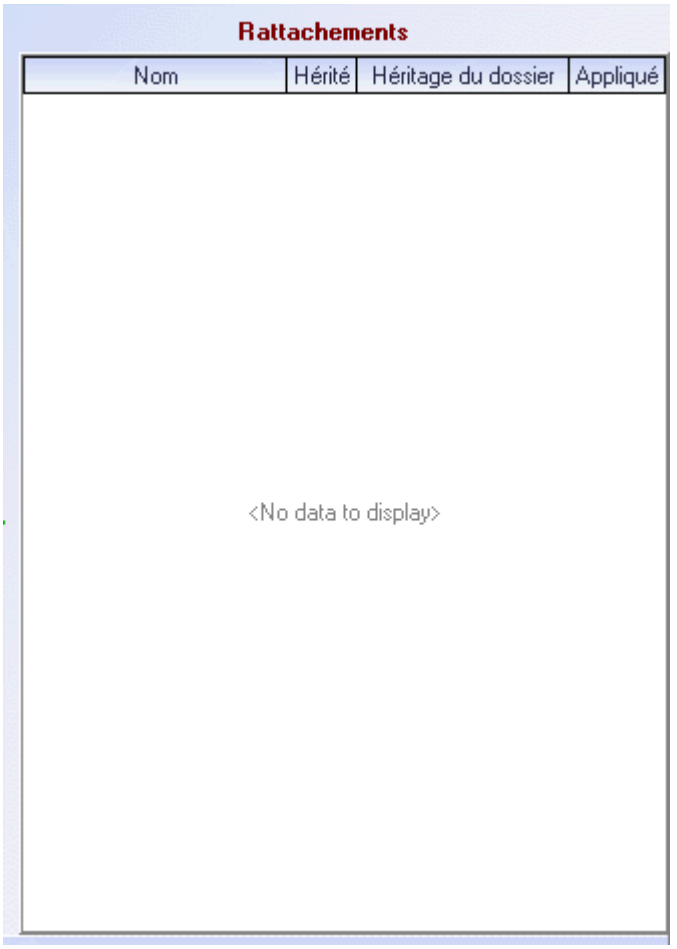
Cocher les cases correspondantes aux options choisies :

- Spécifier les motifs d'alerte dans le fichier d'index (O/N)

- Définir comme format d'export par défaut : lorsque l'on exporte une source directement depuis la fiche principale, depuis l'arbre des sources ou suite à une recherche, c'est le format d'export défini par défaut qui est utilisé.

- N'exporter que les zones scrapées : Lorsque l'on utilise le scraper sur une source, il est possible de choisir de n'exporter que la zone scrapée en cochant cette option.

8.3 Le cadre de droite



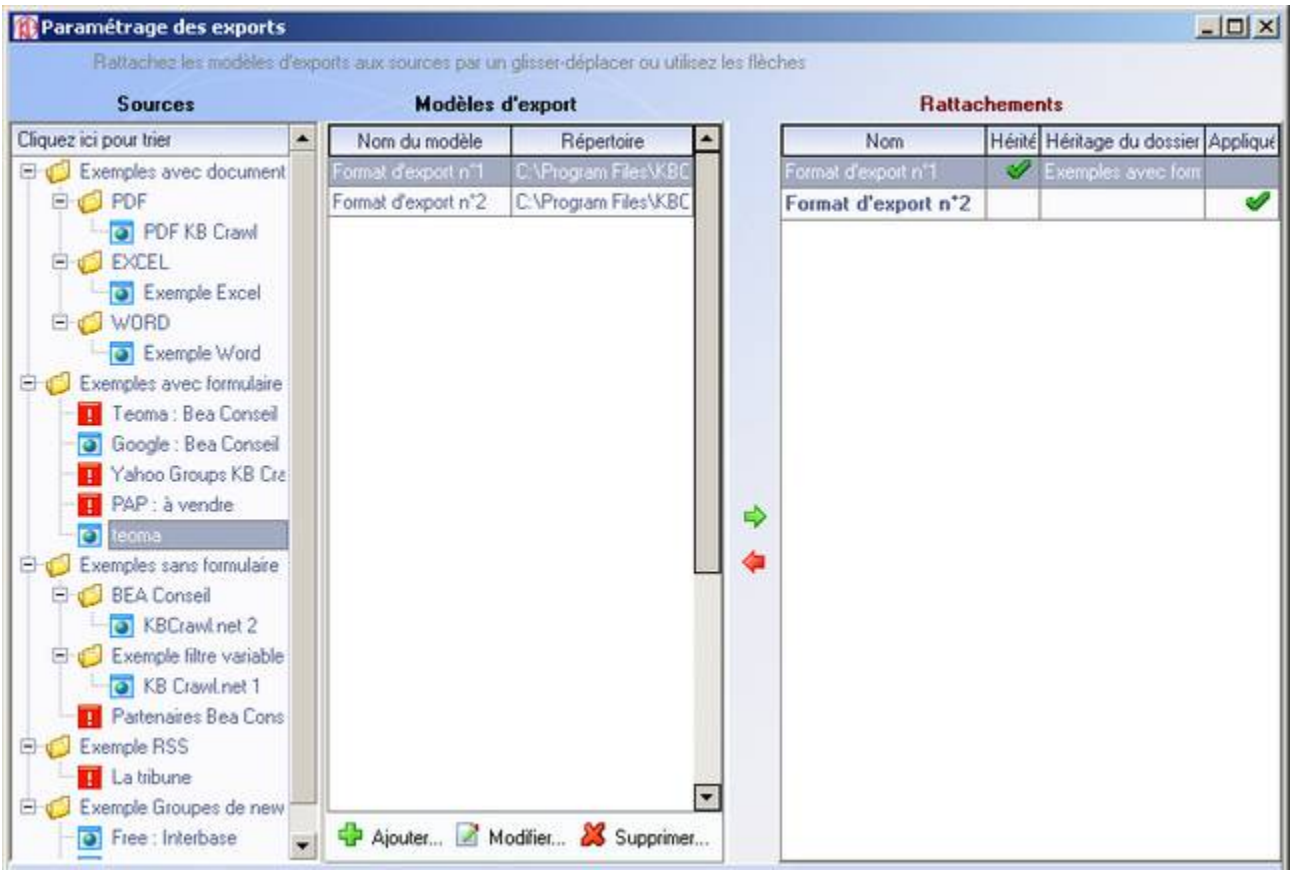
Le cadre de droite montre le format d'export rattaché à une source ou à un dossier donné.

Pour effectuer un rattachement, deux méthodes sont possibles :

- Se positionner sur une source ou un dossier dans le cadre de gauche, sur un modèle d'export dans le cadre du milieu, puis cliquer sur la flèche verte

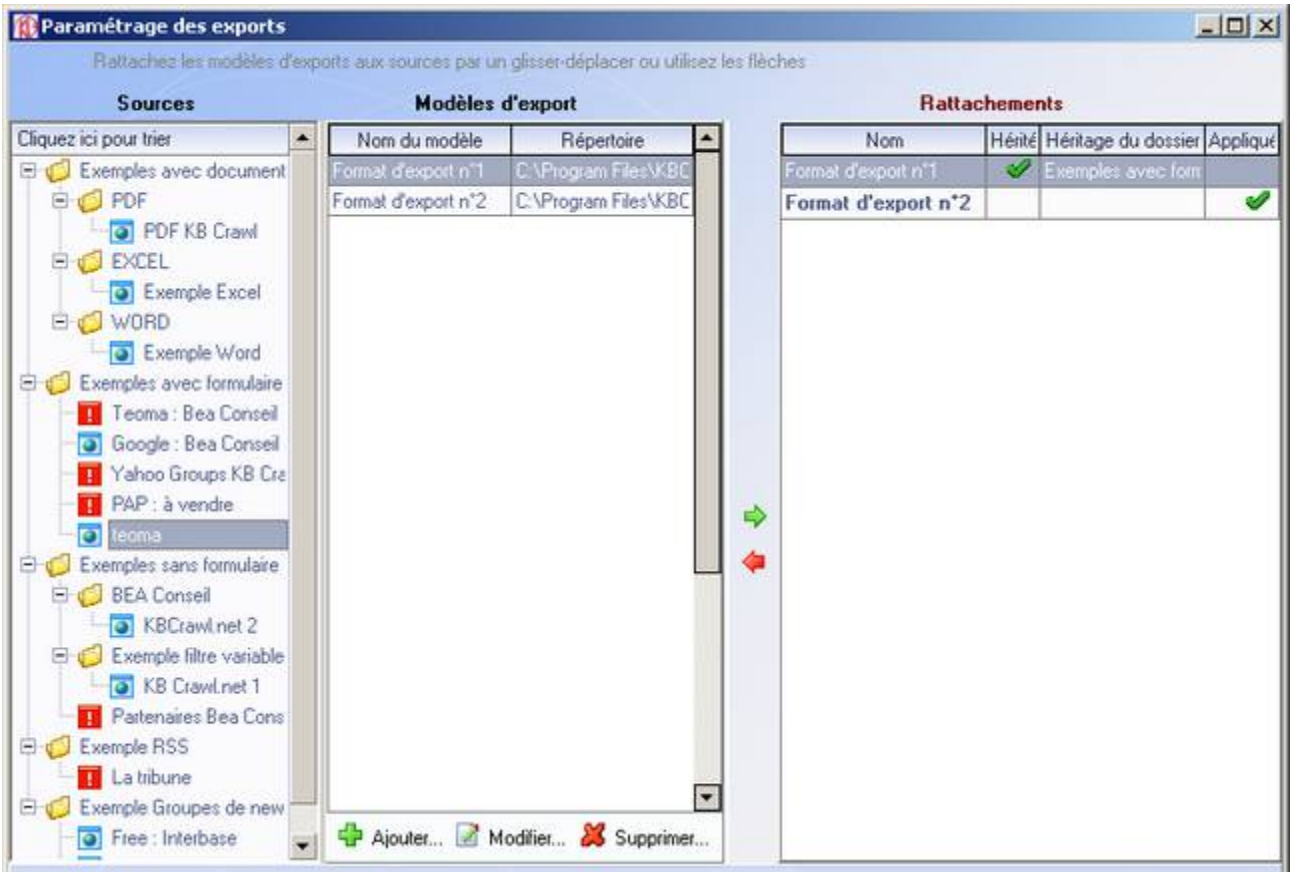
- Effectuer un « glisser-déplacer » depuis la liste des modèles d'export vers une source ou un dossier.

8.4 Héritage des modèles d'export



Si une source est rattachée à un modèle d'export et que cette source est contenue dans un dossier, lui-même rattaché à un modèle d'export, ou encore que plusieurs sous dossiers formant une chaîne de filiation sont rattachés à des modèles d'export différents, la règle qui s'applique est la suivante : c'est le modèle rattaché à l'entité (source ou dossier) du plus bas niveau qui s'applique.

Exemple :



Ici, la source « Teoma » est rattachée au format d'export N°2 alors que le dossier qui la contient est rattaché au format d'export N°1.

C'est dont le modèle d'export rattaché à la source Teoma qui s'applique comme l'indique la grille dans le cadre de droite.

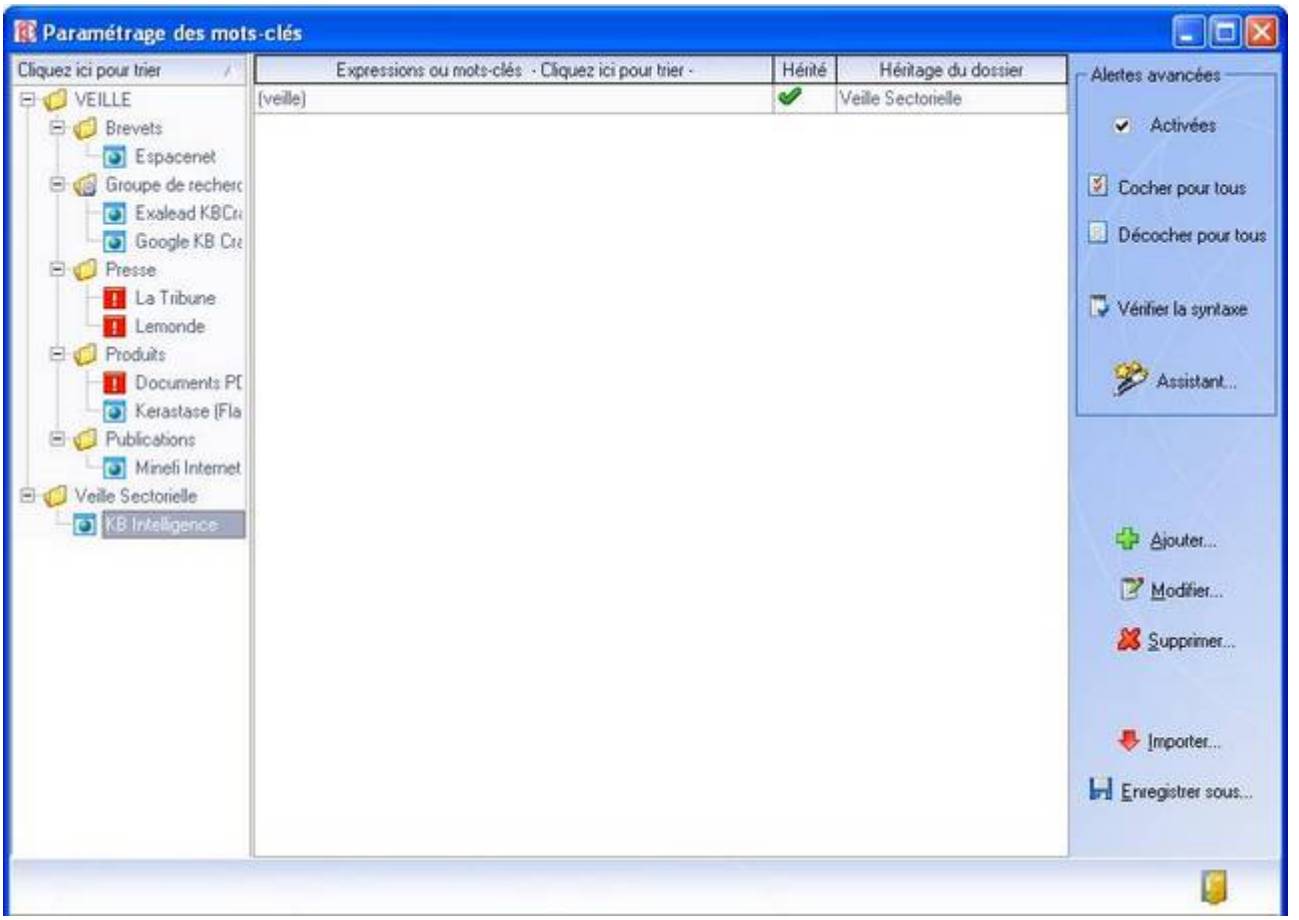
9 Gestion des mots-clés d'alerte

Les mots-clés d'alerte sont affectés à des sources ou à des dossiers.
Chaque source hérite des mots-clés des dossiers et sous dossiers auxquels elles appartiennent.
Lorsque l'on modifie les paramètres d'une source ou bien ceux d'un dossier dans le menu d'options globales, on ne voit que les mots-clés du niveau concerné, ce qui ne permet pas, au final de voir pour une source donnée tous les mots-clés susceptibles de déclencher une alerte.
Le module de gestion des mots-clés d'alerte apporte une solution à ce problème.

On y accède de plusieurs façons :

- en cliquant sur le bouton « Mots-clés » dans la barre de menu principale : 

- directement depuis la source ou le dossier concerné en faisant un clic droit + « mots-clés d'alerte… » (CTRL + K),



Une fenêtre s'ouvre et montre des informations contextuelles à la source ou au dossier sur lequel on était positionné dans la fenêtre principale.

Cette fenêtre possède trois cadres :

- Cadre de gauche

Le cadre de gauche présente la liste des sources et des dossiers sous forme arborescente, similaire en tous points à celui présent dans la fenêtre principale.
On peut ainsi sélectionner la source ou le dossier pour lesquels on souhaite paramétrer les mots-clés d'alerte.

- Cadre central

Le cadre central présente la liste des mots-clés contextuels à l'entité sélectionnée dans le cadre de gauche, triée par ordre alphabétique.
L'entité peut être un dossier ou bien une source. Quel que soit le cas, on observe deux types d'expressions ou mots-clés : ceux qui sont directement rattachés à l'entité, et ceux qui sont hérités d'une entité contenant directement ou indirectement celle-ci.

La grille du cadre central contient un champ « hérité » prévu à cet effet. Si l'expression d'alerte est héritée, le dossier propriétaire de cette expression est notifiée dans la colonne « Héritage du dossier ».

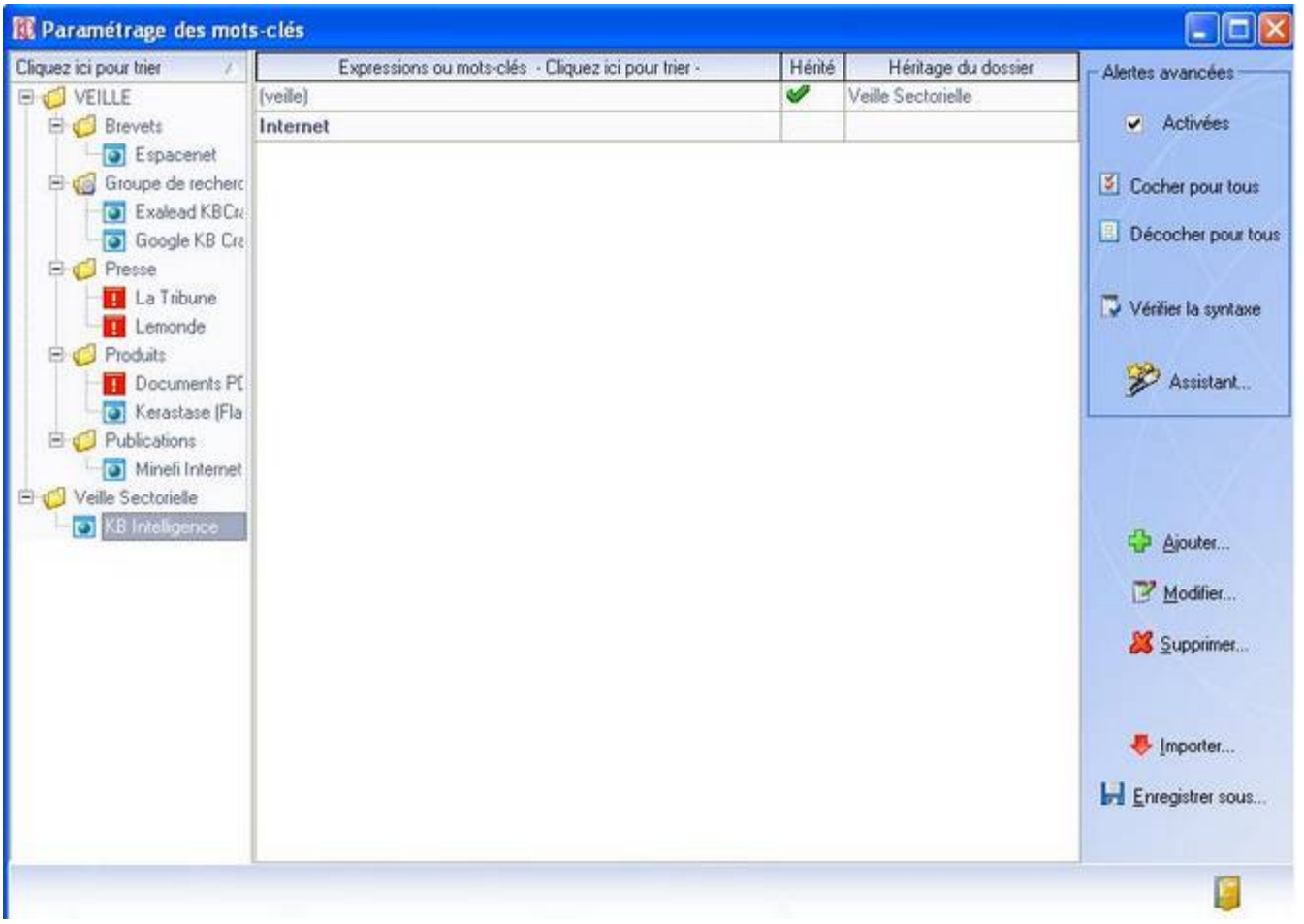


Figure 54 : Mots-clés d'une source

- Cadre de droite

Le cadre de droite sert principalement à ajouter, modifier ou supprimer des mots-clés ou expressions d'alertes une fois que l'on a sélectionné l'entité qui en sera propriétaire dans le cadre de gauche.
Pour cela utiliser les boutons portant les libellés correspondants.

On peut également importer une liste de mots-clés depuis un fichier texte dans lequel chaque mot ou expression d'alerte est séparé par un retour chariot.
Pour cela, utiliser le bouton « importer… »

De plus, le bouton « enregistrer sous… » permet d'exporter la liste des mots-clés ou expressions d'alerte au format texte.

Le cadre de droite permet également de spécifier le type d'élément d'alerte sélectionné dans le cadre central.
Pour activer le mode « alerte avancée » pour une expression donnée, cocher la case « activées ».
On peut activer le mode « alerte avancée » pour l'ensemble des expressions listées dans le cadre central, en cliquant sur le bouton « cocher pour tous » et inversement en cliquant sur le bouton « décocher pour tous ».

Lorsque le mode « alerte avancée » est sélectionné, lors du crawl, c'est le moteur de recherche de KB Crawl qui sera interrogé avec une expression résultante de la liste des expressions saisies :

Liste d'expressions :
<expr A>
<expr B>
<expr C>

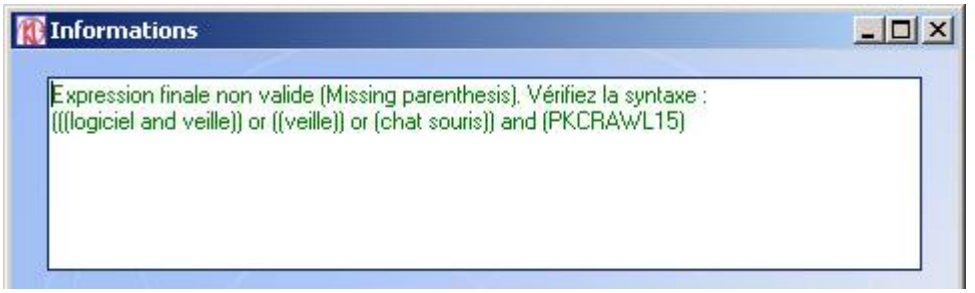
Expression finale : <expr A> or <expr B> or <expr C>

Pour vérifier que la syntaxe de cette expression finale est valide, et que donc, le mécanisme d'alerte avancé s'effectuera sans anomalie, il suffit de cliquer sur le bouton « Vérifier la syntaxe ».

Un message d'information apparaît alors :

Exemple d'erreur :

On clique sur « vérifier la syntaxe »…



On voit ici l'expression finale. En effet, les termes « chat » et « souris » devraient être séparés par des opérateurs logiques.
Le terme PKCRAWL15 est un identifiant qui permet de filtrer la recherche sur la source, il est placé automatiquement dans l'expression finale, et n'est pas à prendre en compte.

Il est possible de créer des requêtes avancées afin d'affiner sa surveillance et de n'être alerté que dans certains cas précis. Les opérateurs booléens permettent de le faire. Les opérateurs booléens utilisables sont les suivants :

- and : à utiliser lorsque l'on souhaite être alerté si tous les mots de la requête figurent sur les pages surveillées
- « » : à utiliser lorsque l'on souhaite être alerté sur l'apparition d'une expression exacte ; celle-ci devra être écrite entre guillemets
- or : à utiliser lorsque l'on souhaite être alerté si un ou plusieurs mots de la requête figurent sur les pages surveillées
- not : à utiliser lorsque l'on ne souhaite pas être alerté si le mot figure sur l'une des pages surveillées (élimination du bruit)
- Near : à utiliser lorsque l'on souhaite être alerté si les deux mots spécifiés soient éloignés de huit mots au maximum

Ces opérateurs peuvent être utilisés manuellement, mais aussi grâce à l'assistant, qui permet de renseigner les champs sans avoir à taper les opérateurs : chaque champ correspond à un opérateur.

- Le champ « Tous les mots suivants » : correspond à l'opérateur and
- Le champ « L'expression exacte » : correspond aux guillemets
- Le champ « Au moins l'un des mots suivants » : correspond à l'opérateur or
- Le champ « Aucun des mots suivants » : correspond à l'opérateur not

- Les champs « les deux mots suivants éloignés de 8 mots au maximum » : correspondent à l'opérateur near.

10 Fonction recherche

10.1 Fonctionnement général

Le moteur de recherche permet de retrouver à l'intérieur de la base de données de KB Crawl, toute page dont le contenu a été indexé par le moteur de recherche KB Crawl.
On entend par contenu le texte de la page lisible depuis un navigateur.

Pour accéder à ce module, cliquer sur le bouton « recherche » de la barre d'outils générale 

Si un ou plusieurs contenus répondent aux critères de la requête soumise au moteur de recherche, les pages correspondantes sont récupérées du module d'archives pour être présentées avec les mots correspondant aux résultats de la recherche.

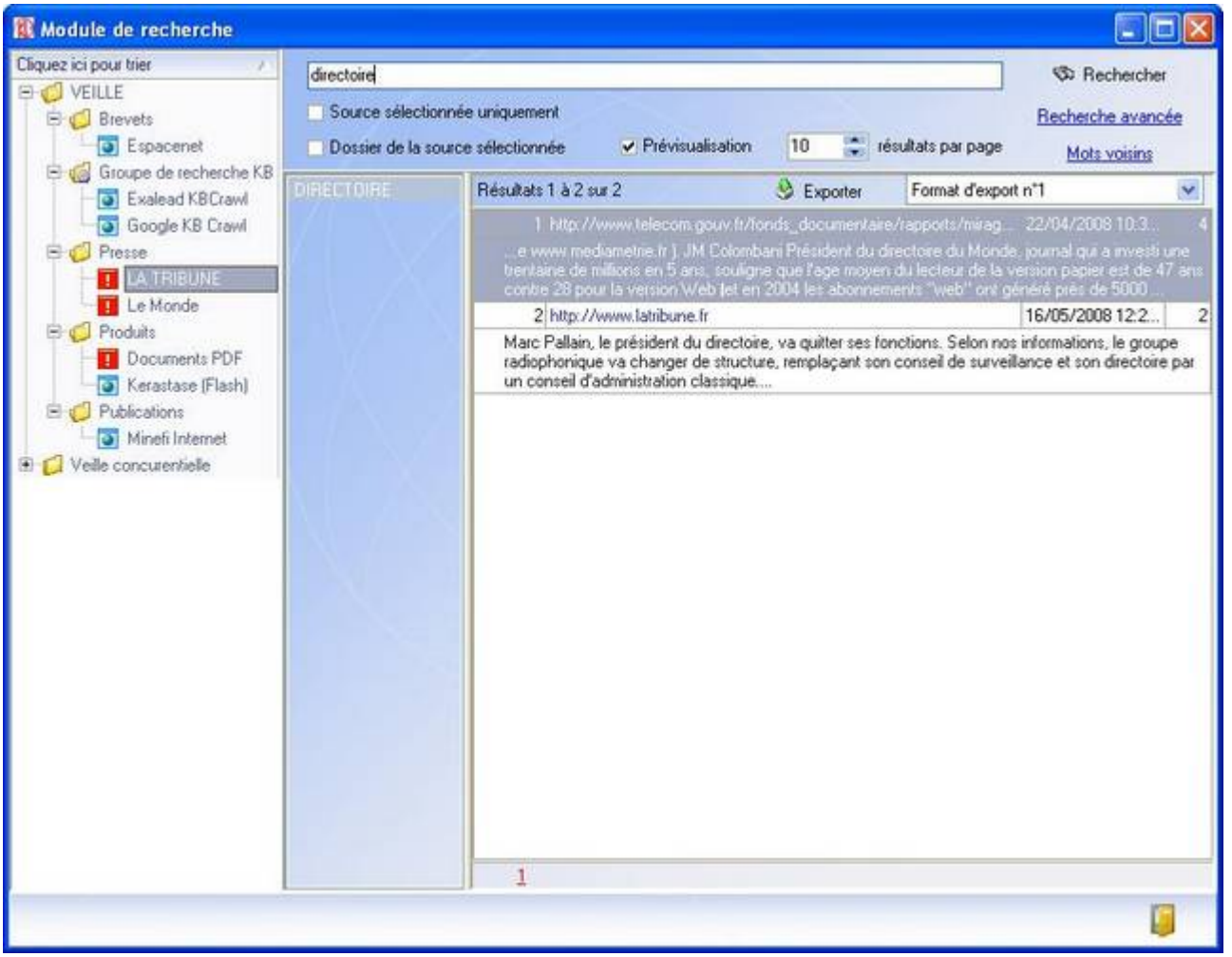


Figure 55 : Résultats d'une recherche.

10.2 Ergonomie générale

A gauche de la partie « recherche » de l'écran, on retrouve la liste des sources et des dossiers sous forme arborescente, similaire en tous points à celui présent dans la fenêtre principale.

La partie recherche se décompose ensuite en quatre parties :

10.2.1 Cadre du haut

Le panneau du haut contient différents éléments :

- La zone de saisie
  C'est dans cette zone qu'il faut saisir la requête de recherche.

- L'option « Source sélectionnée uniquement »
  Permet de restreindre la recherche à la source sélectionnée.

- L'option « Dossier de la source sélectionnée »
  Permet de restreindre la recherche aux dossiers de la source sélectionnée.

- L'option « prévisualisation »
  Permet d'activer la prévisualisation des contenus des pages trouvées par le moteur de recherche afin éventuellement d'optimiser le temps de présentation des résultats et aussi l'espace occupé par chaque enregistrement résultat.

- Le nombre de résultats par page
  Permet de paramétrer le nombre d'enregistrements résultats de la recherche par page.

- Le bouton « Rechercher »
  Lance la recherche par rapport à la requête saisie dans la zone prévue à cet effet.

- Le lien « Recherche avancée »
  Fait surgir l'assistant pour les requêtes avancées.

- Le lien « Mots voisins »
  Lance une recherche avec des mots « voisins » de celui utilisé précédemment pour la recherche.

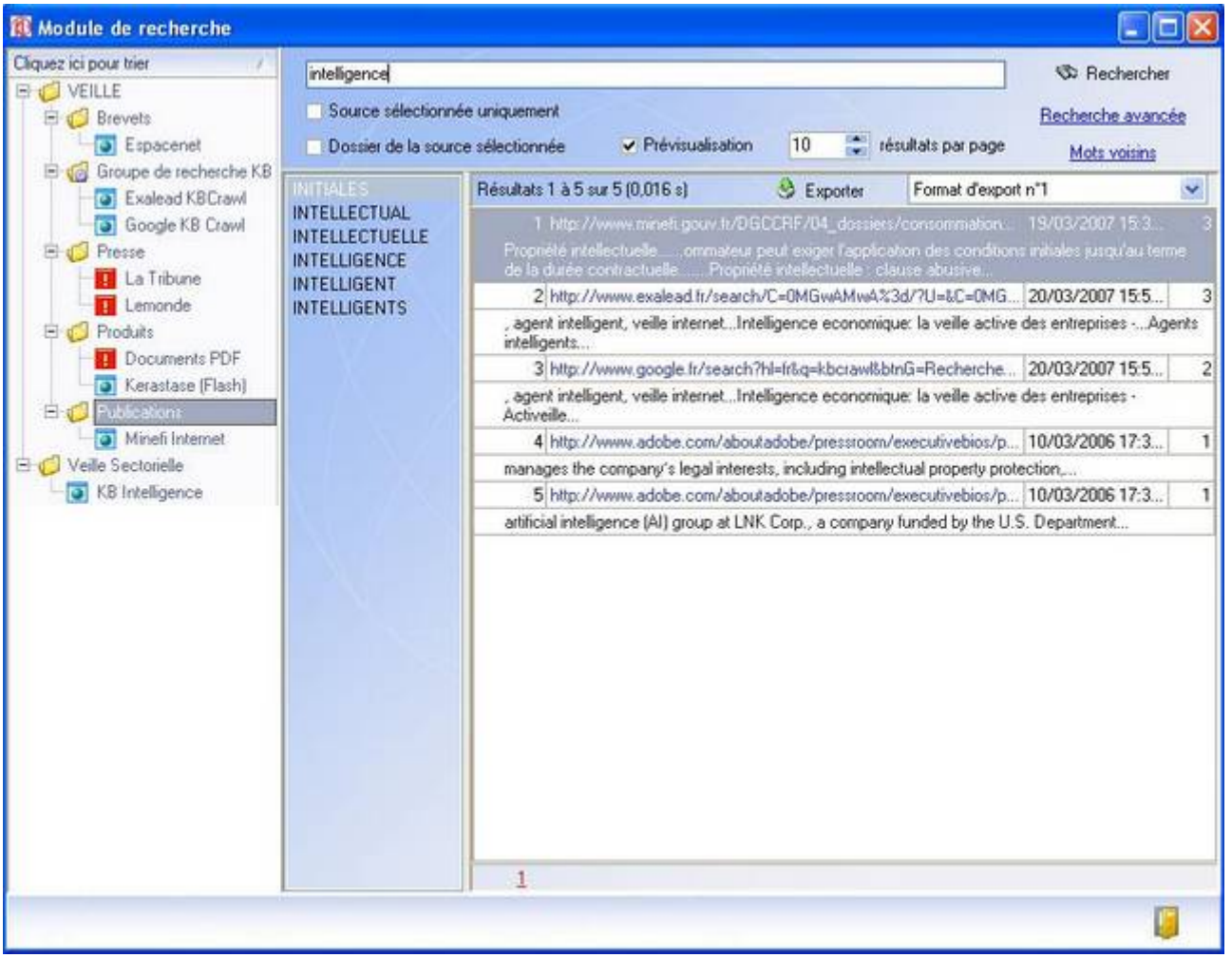


Figure 56 : Mots voisins.

10.2.2 Cadre du bas

Permet d'accéder aux différentes pages de résultats.
Si le nombre de pages résultats est supérieur à 10, le bouton « >> » permet d'accéder aux dix pages suivantes et le bouton « << » aux dix précédentes.

Les différents cas de figure :

Figure 57 : Les 10 premières pages de résultats.

Figure 58 : Les 10 pages de résultats suivantes.

Figure 59 : Les dernières pages de résultats.

10.2.3 Cadre de gauche

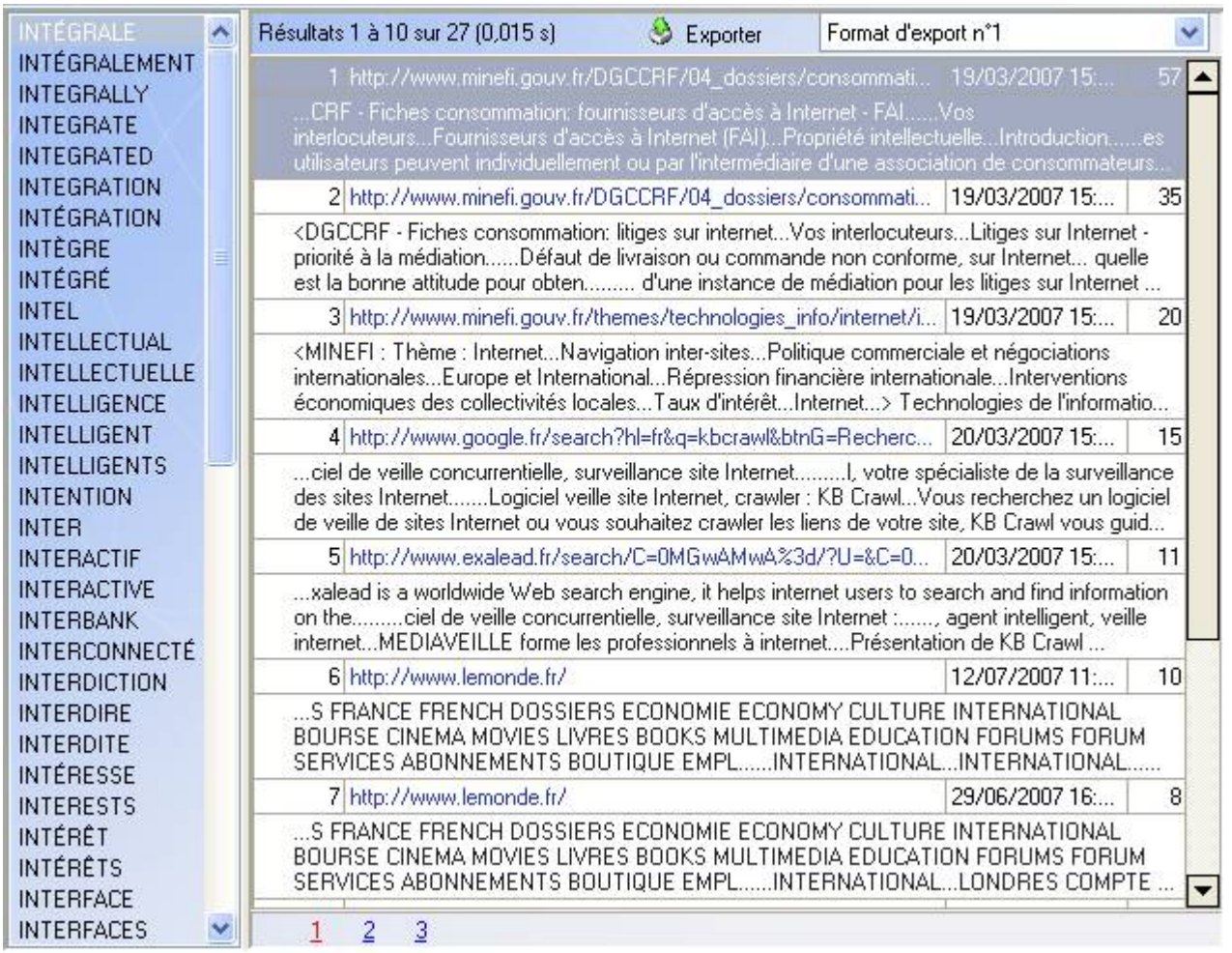


Figure 60 : cadre de gauche

Le cadre de gauche affiche la liste des mots correspondant aux résultats de la recherche.
Ici par exemple, on a utilisé la troncature int* pour effectuer une recherche. Le moteur de recherche KB Crawl renvoie alors la liste des mots correspondant à cette recherche.

En cliquant sur un des mots de la liste, on restreint le périmètre des résultats de recherche au mot sélectionné.

10.2.4 Cadre de droite

Le cadre de droite contient une grille qui affiche la liste des enregistrements résultats de la recherche.
Chaque enregistrement contient plusieurs colonnes d'informations :

- l'URL correspondant à la page,
- la date du crawl,
- le classement de la page au sein des pages de résultats,
- Une prévisualisation des extraits du contenu qui contient les mots-clés de la recherche.

Lorsque l'on double-clique sur l'un des enregistrements de la grille, le navigateur par défaut s'ouvre pour monter le document correspondant contenant les mots-clés de recherche surlignés :

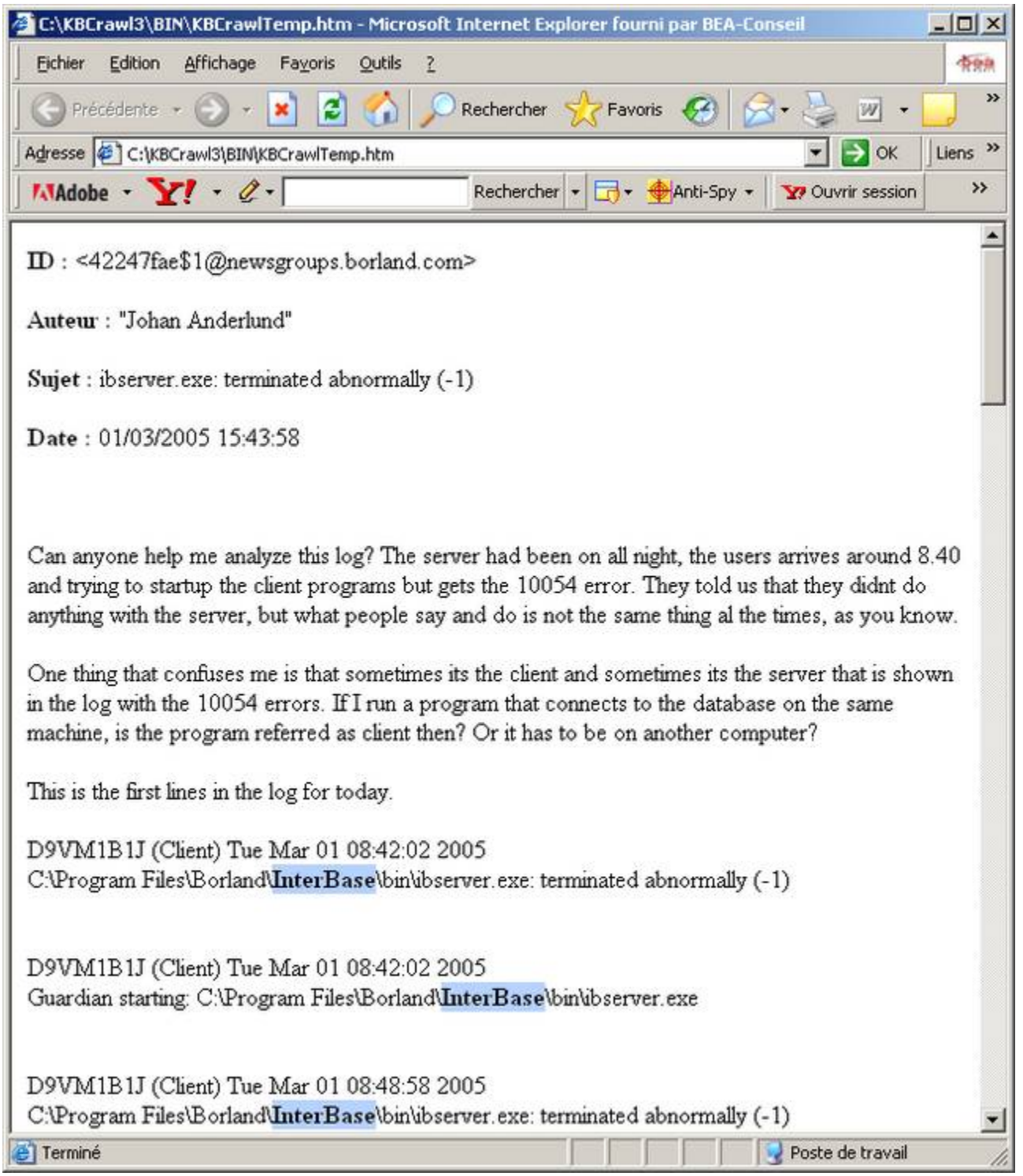


Figure 61 : Visualisation d'une page résultat d'une recherche dans le browser.

Le cadre de droite présente une partie haute qui affiche le nombre de résultats obtenus par la recherche avec le temps mis par le moteur de recherche pour trouver les résultats. A ce temps

doit être ajouté le temps pris pour renseigner la grille de ces enregistrements et créer les parties de texte de prévisualisation.

Cette partie propose également un bouton « exporter » qui permet de placer toutes les pages résultat affichées dans la grille dans un répertoire choisi avec une page d'index (se reporter au chapitre « export »).
L'export se fait alors au format d'export choisi dans la liste déroulante juste à côté du bouton « export ». KB Crawl crée alors un répertoire à partir du répertoire défini dans le format d'export, nommé ainsi : recherche_jjmmaaaa hh_mm_ss.

10.3 Effectuer une recherche

Pour effectuer une recherche, il faut se placer dans l'onglet « Recherche ».
Celle-ci peut s'effectuer sur la source sélectionnée ou le dossier sélectionné selon que l'on coche ou non l'une des cases « Source sélectionnée uniquement» ou « Dossier de la source sélectionnée ». Si l'on ne choisit aucune de ces deux options, la recherche s'effectue sur la totalité des sources contenues dans la base de données.
La recherche d'informations se fait par des requêtes, au même titre que n'importe quel moteur de recherche sur Internet : Ces requêtes sont composées d'un ou plusieurs mots-clés ou expressions, éventuellement séparés par des opérateurs booléens (« and », « or », « not », etc.).

Nota :
- Le moteur de recherche n'est pas sensible à la casse.
- Les opérateurs « and » et « not » remplacent respectivement « + » et « - » utilisés dans les versions précédentes de KB Crawl.

Pour effectuer une requête, saisir une requête de recherche dans la zone de saisie :

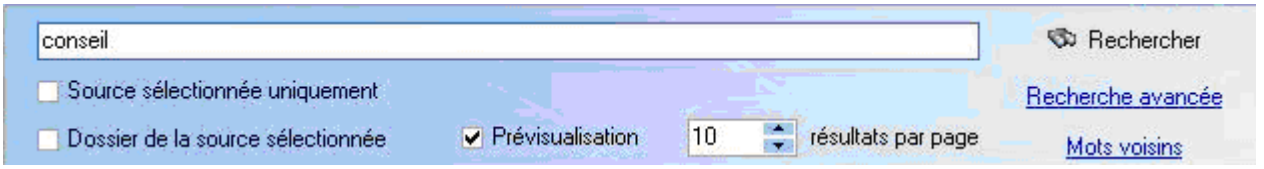


Figure 62 : Fonction de recherche.

10.3.1 Requête simple

Ici, on cherche les pages qui contiennent le mot « directoire ».
Le moteur de recherche ne tient jamais compte des minuscules/majuscules, par contre il tient compte des accents.

Exemple 1 : Recherche avec le mot « conseil ».

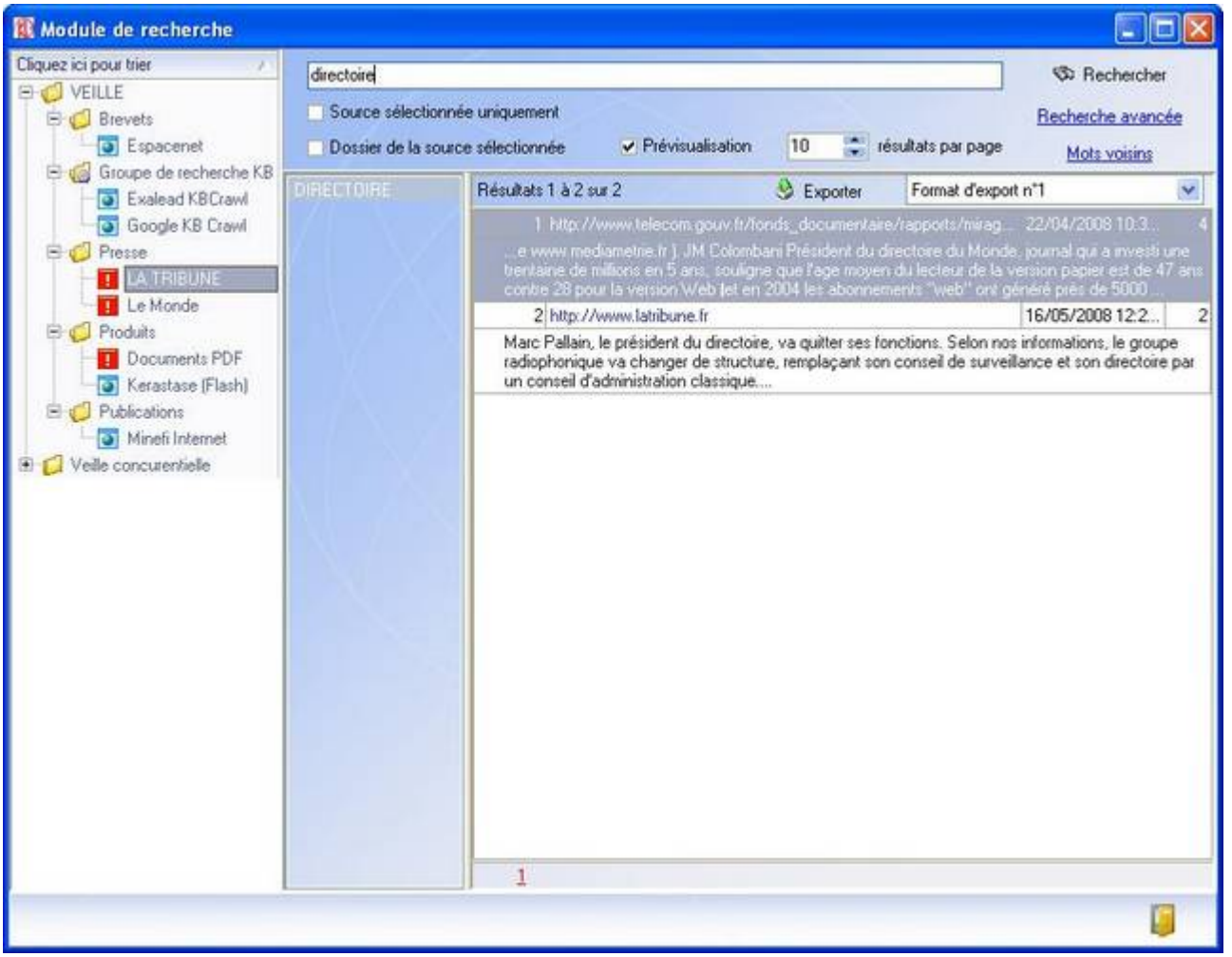


Figure 63 : Fonction de recherche (exemple 1).

10.3.2 Requête avec booléens

Exemple 2 : veille and internet

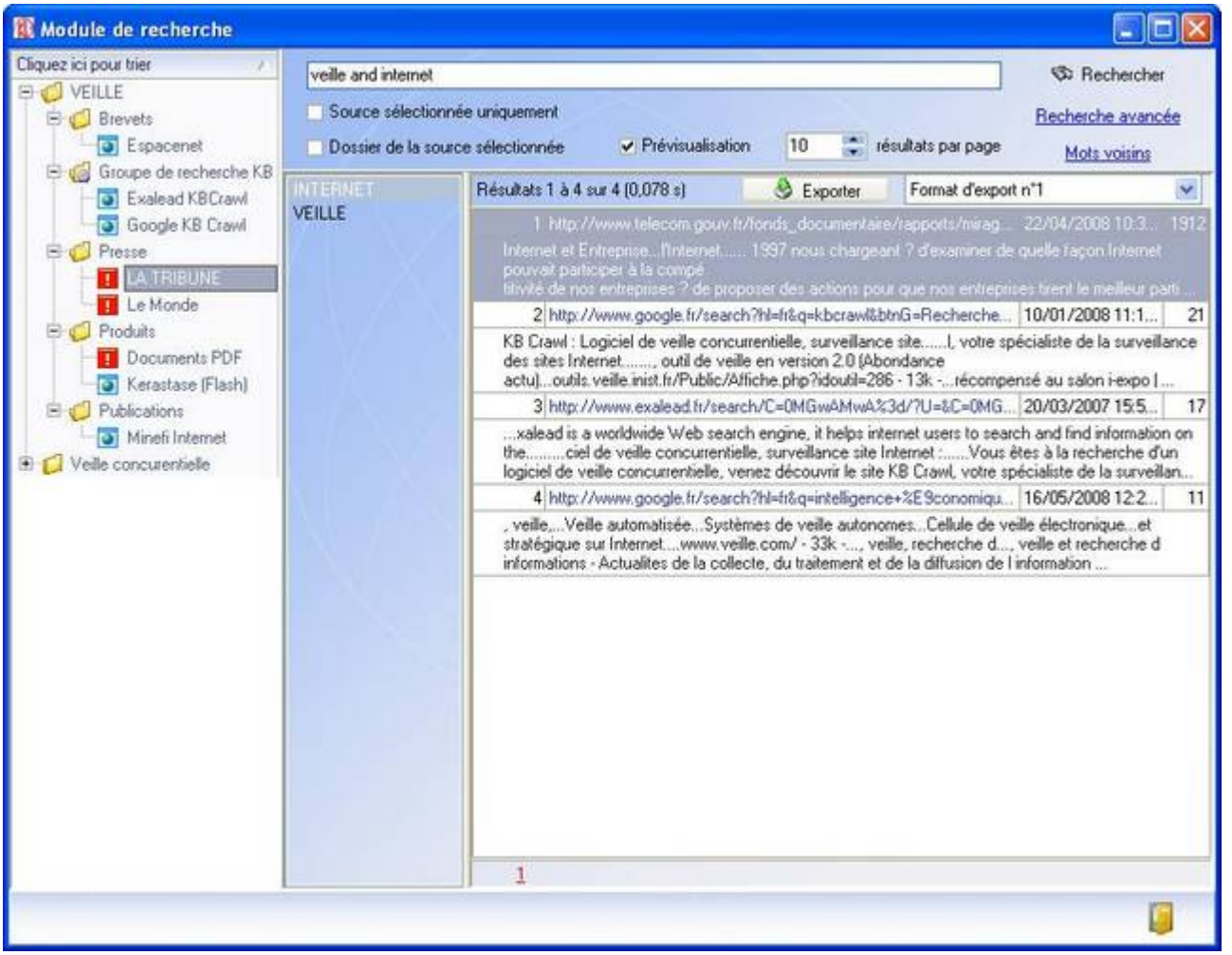


Figure 64 : Fonction de recherche (exemple 2).

Lorsque deux mots-clés de recherche sont séparés par un « and », cela signifie que les deux mots-clés doivent figurer dans le contenu des pages renvoyées.

Exemple 3 : (veille and internet) and not (exalead)

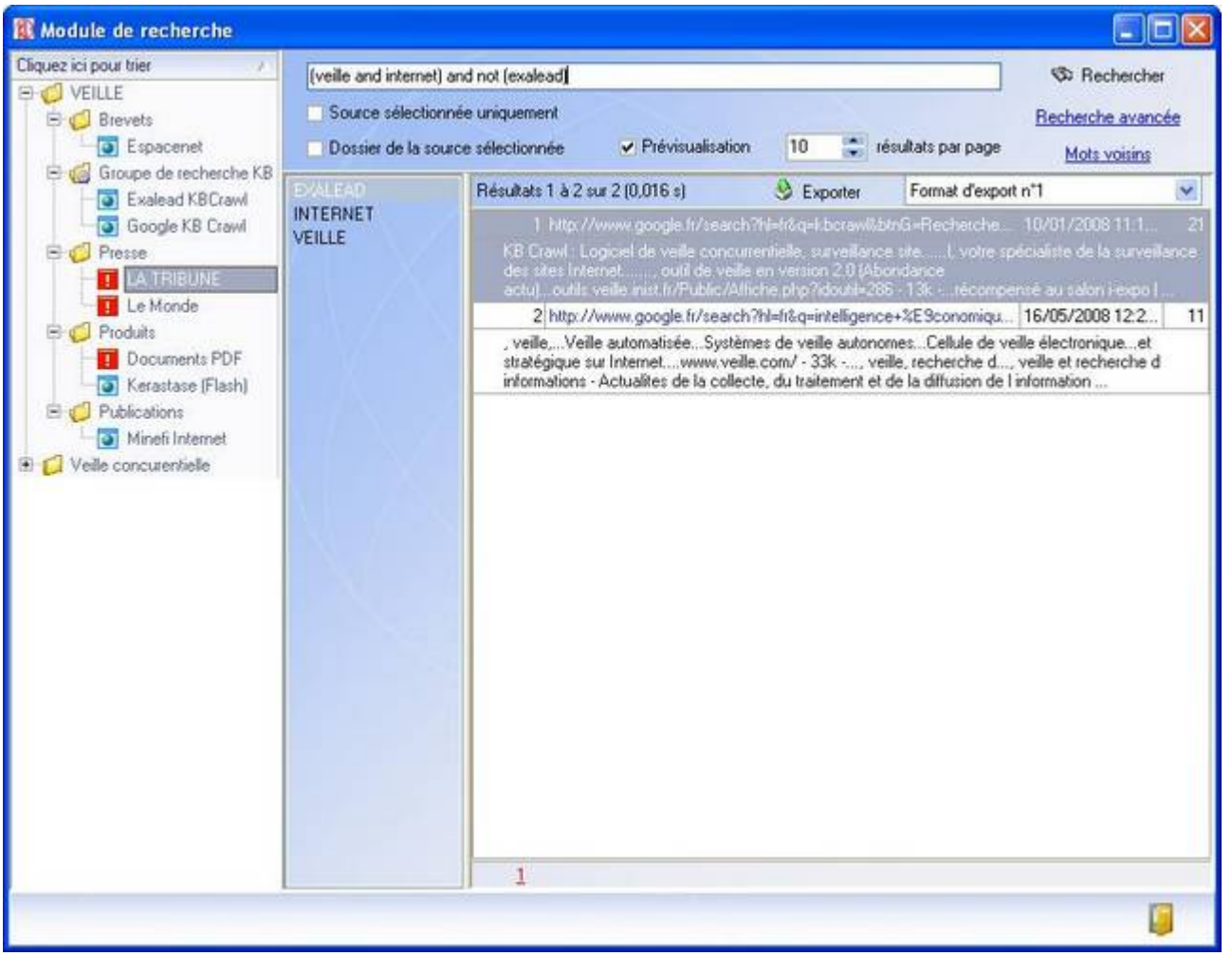


Figure 65 : Fonction de recherche (exemple 3).

Ici, il ne reste que deux résultats car on n'a pas souhaité que le mot « exalead» soit présent dans les pages résultats.

Exemple 4 : (veille) and ("recherche avancée")

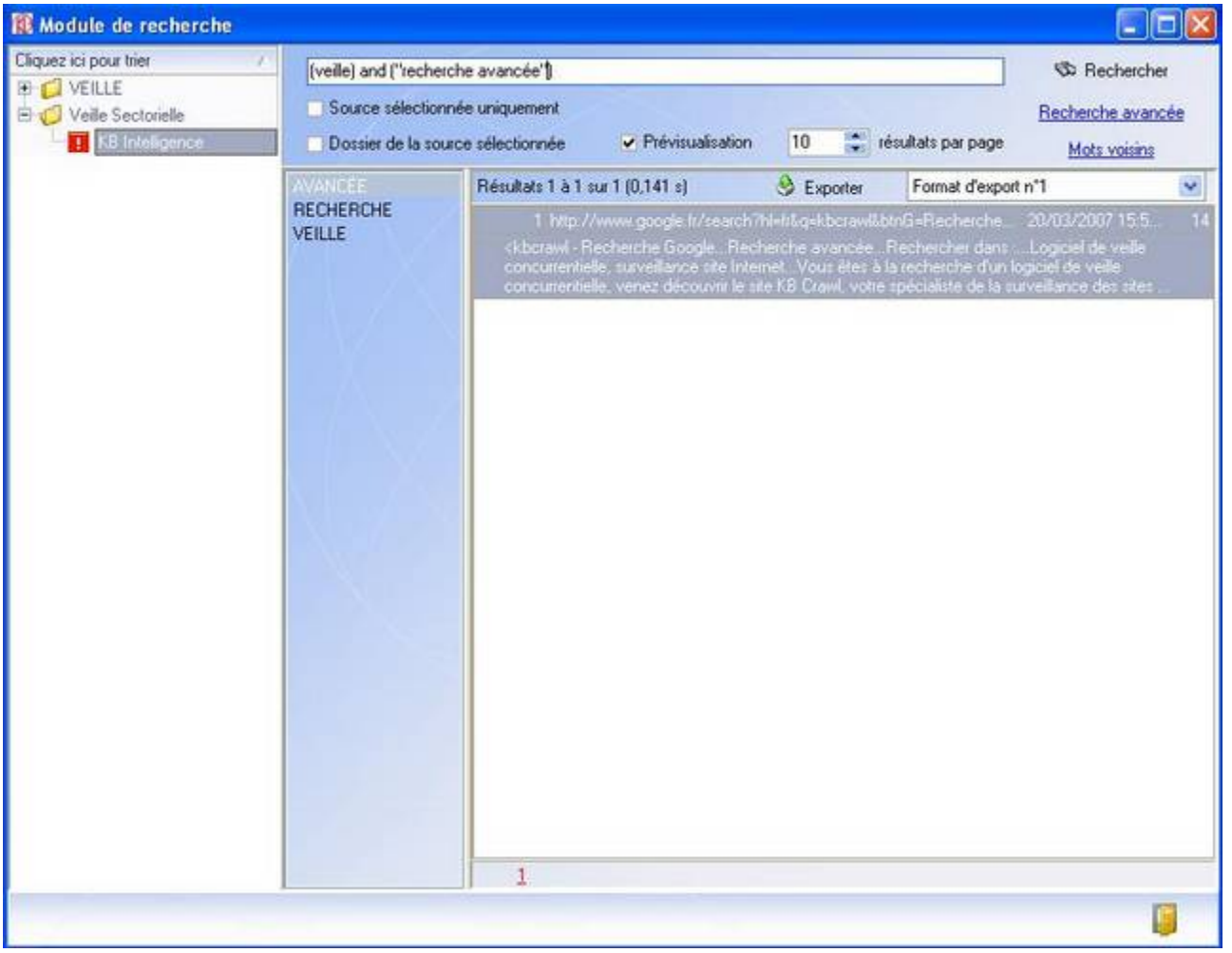


Figure 66 : Fonction de recherche (exemple 4).

Ici, on souhaite que la phrase « recherche avancée » soit contenue dans la page.
Il faut pour cela encadrer la phrase avec des guillemets. Les guillemets sont nécessaires pour tout élément d'une expression contenant plus d'un mot. Ici, « recherche avancée » est le deuxième élément de l'expression et contient deux mots, d'où la nécessité des guillemets.

Exemple 5 : veille or internet

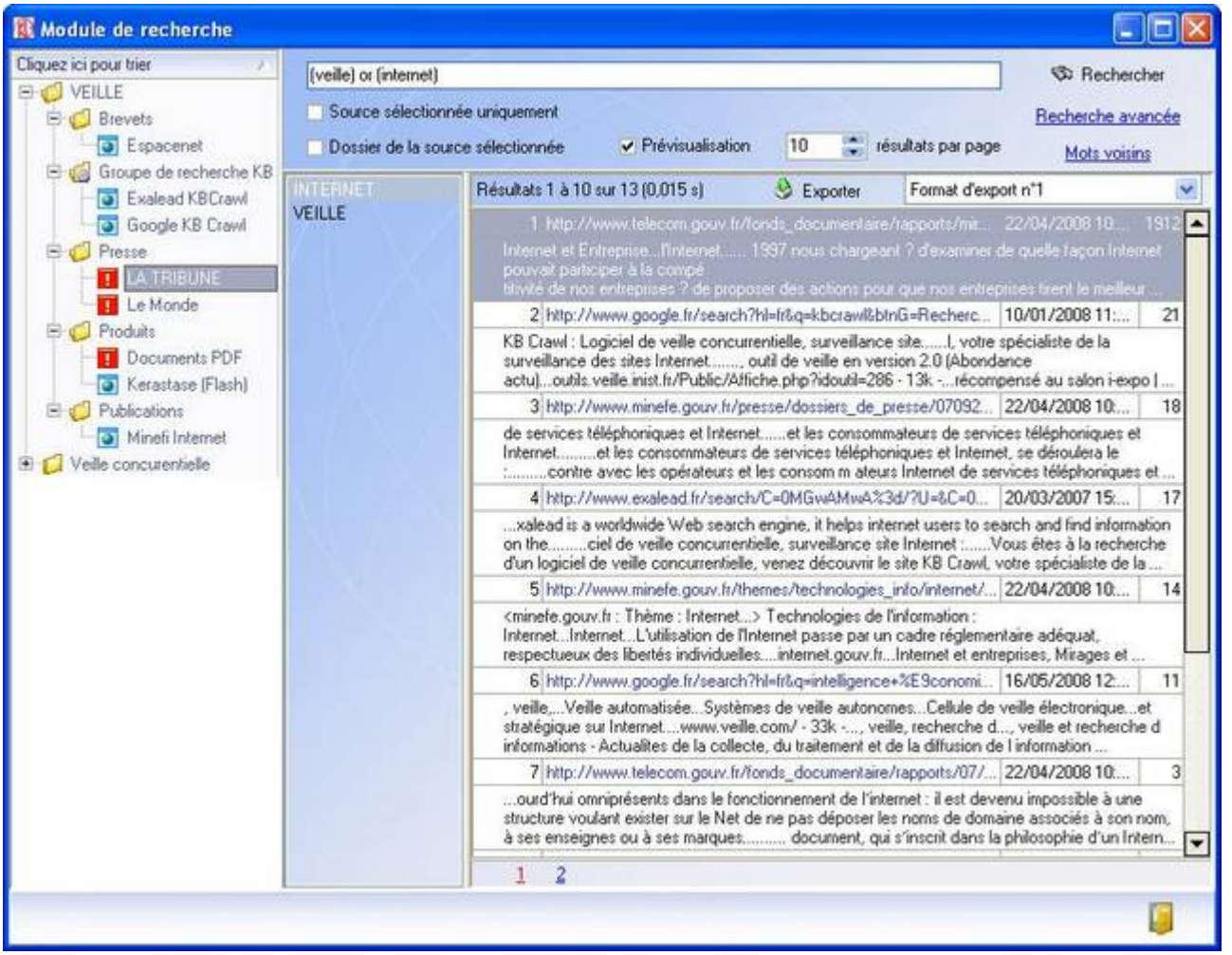


Figure 67 : Fonction de recherche (exemple 5).

Ici, les mots-clés de la requête ne sont pas séparés par des opérateurs, ce qui signifie que l'un ou l'autre des deux mots-clés doit se trouver dans la page résultat.

Attention ci-dessus, les mots sont séparés par un or dans le screenshot

Exemple 6 : crawl or (chasse and performante)

On peut également réaliser des combinaisons plus complexes : ici, on veut que CRAWL soit présent dans la page, ou que [CHASSE et PERFORMANTE] soient présents dans la page.

Une fois la requête envoyée, KB Crawl l'analyse pour l'interpréter : il regroupe les opérateurs de même signe.

Exemple 7 : ((chasse and performante and loi) or (crawl or espèce)) and not(treuil)

Interprétée, cette requête signifie : On veut voir les pages avec [CHASSE et PERFORMANTE et LOI] ou [CRAWL ou ESPECE] mais pas TREUIL.

10.3.3 Gestion des troncatures

Pour éviter d'avoir à saisir une requête contenant différents termes ayant la même racine, on peut utiliser la troncature.

Ici, un exemple avec une troncature illimitée avec la racine info :

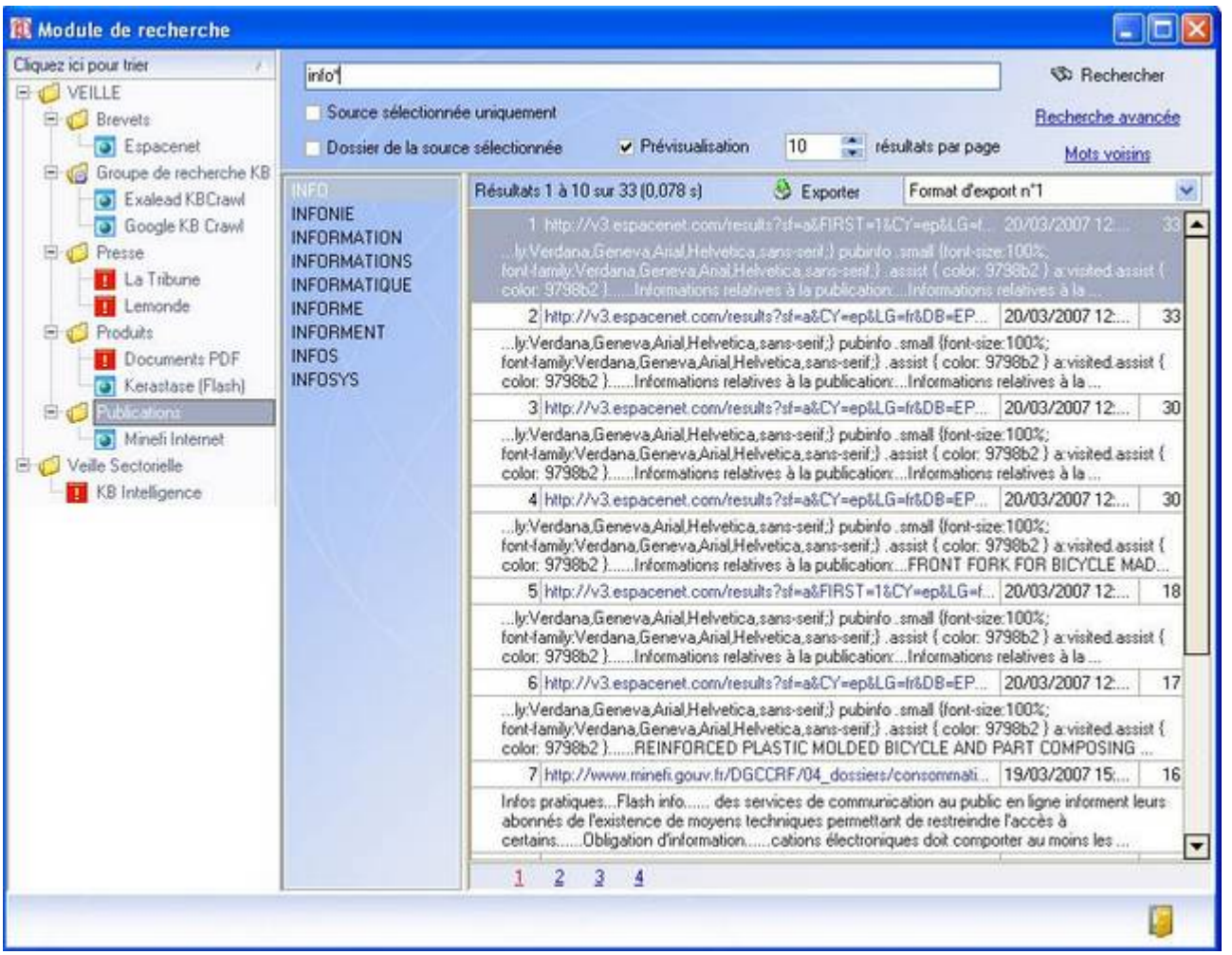


Figure 68 : Troncature (exemple 1).

Le moteur de recherche de KB Crawl renvoie non seulement les pages résultant de la recherche mais également tous les termes trouvés répondant à la troncature.
En cliquant sur un des termes trouvés présentés dans la colonne de gauche, on affine la recherche en précisant qu'elle se fait sur le terme exact sélectionné :

La troncature peut s'appliquer autant de fois que l'on souhaite pour un seul terme et à n'importe quel endroit du terme :

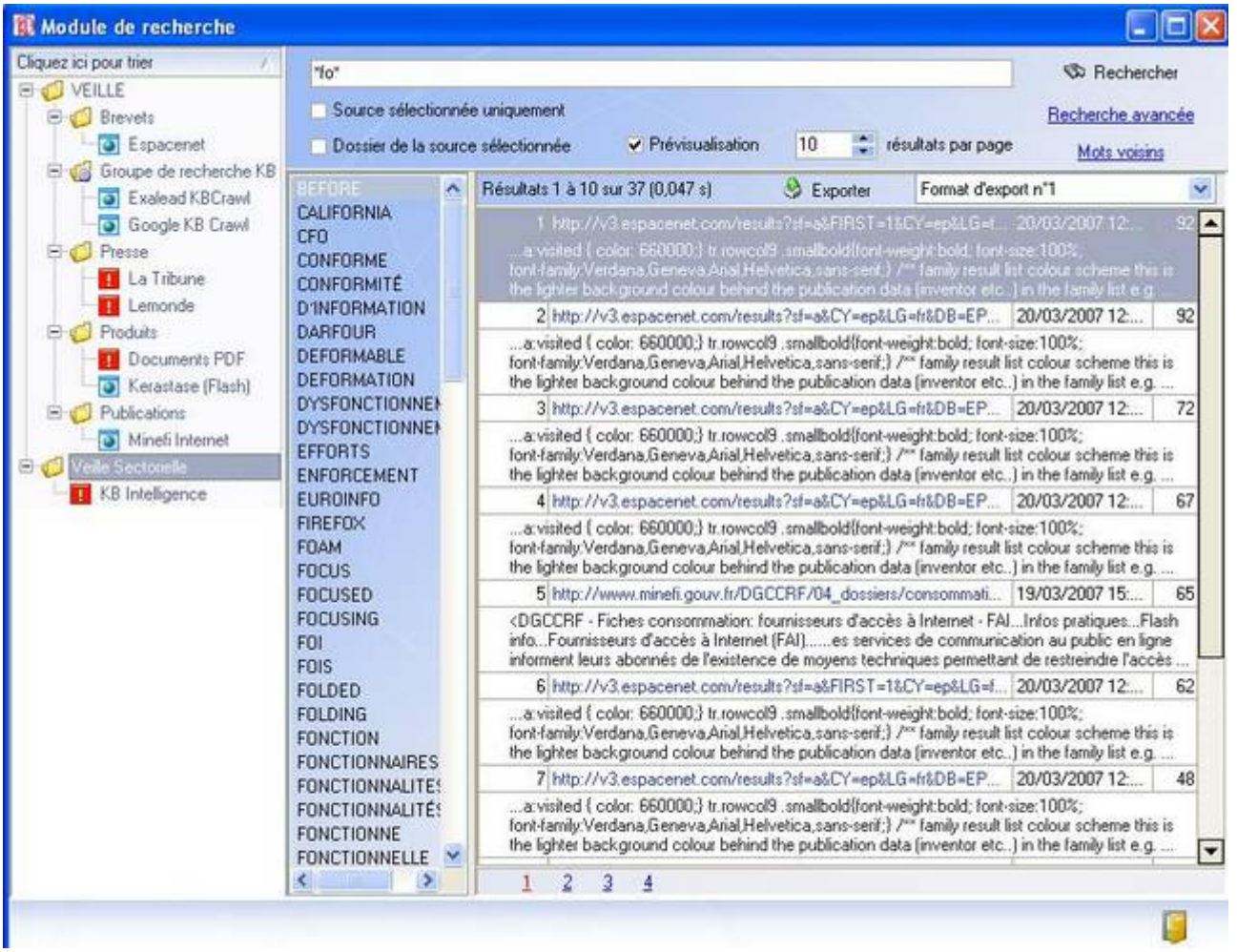


Figure 69 : Troncature (exemple 2) : *fo*

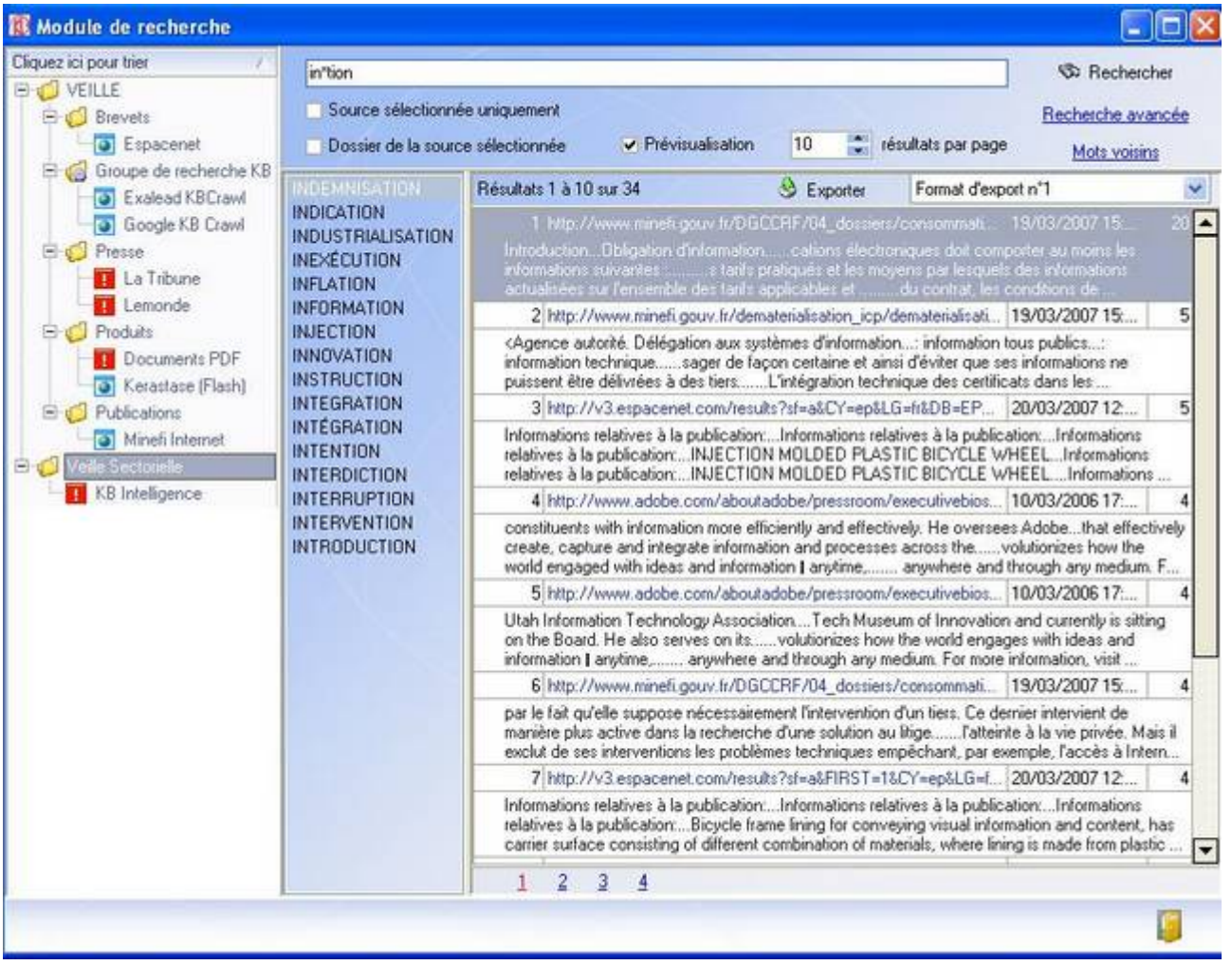


Figure 70 : Troncature (exemple 4) : in*tion.

10.3.4 Gestion des masques

Le masque est un symbole utilisé dans les requêtes de recherche pour remplacer une ou plusieurs lettres à l'intérieur d'un mot. Il s'agit d'un point d'interrogation (?). Il est utile lorsque l'orthographe de certains mots ne varie que très légèrement.

Nota :
Chaque point d'interrogation ne remplacera qu'un seul et unique caractère (il ne remplacera donc pas un espace) ; il est cependant possible d'en utiliser plusieurs dans le même mot.

Quelques exemples d'utilisation de masques :

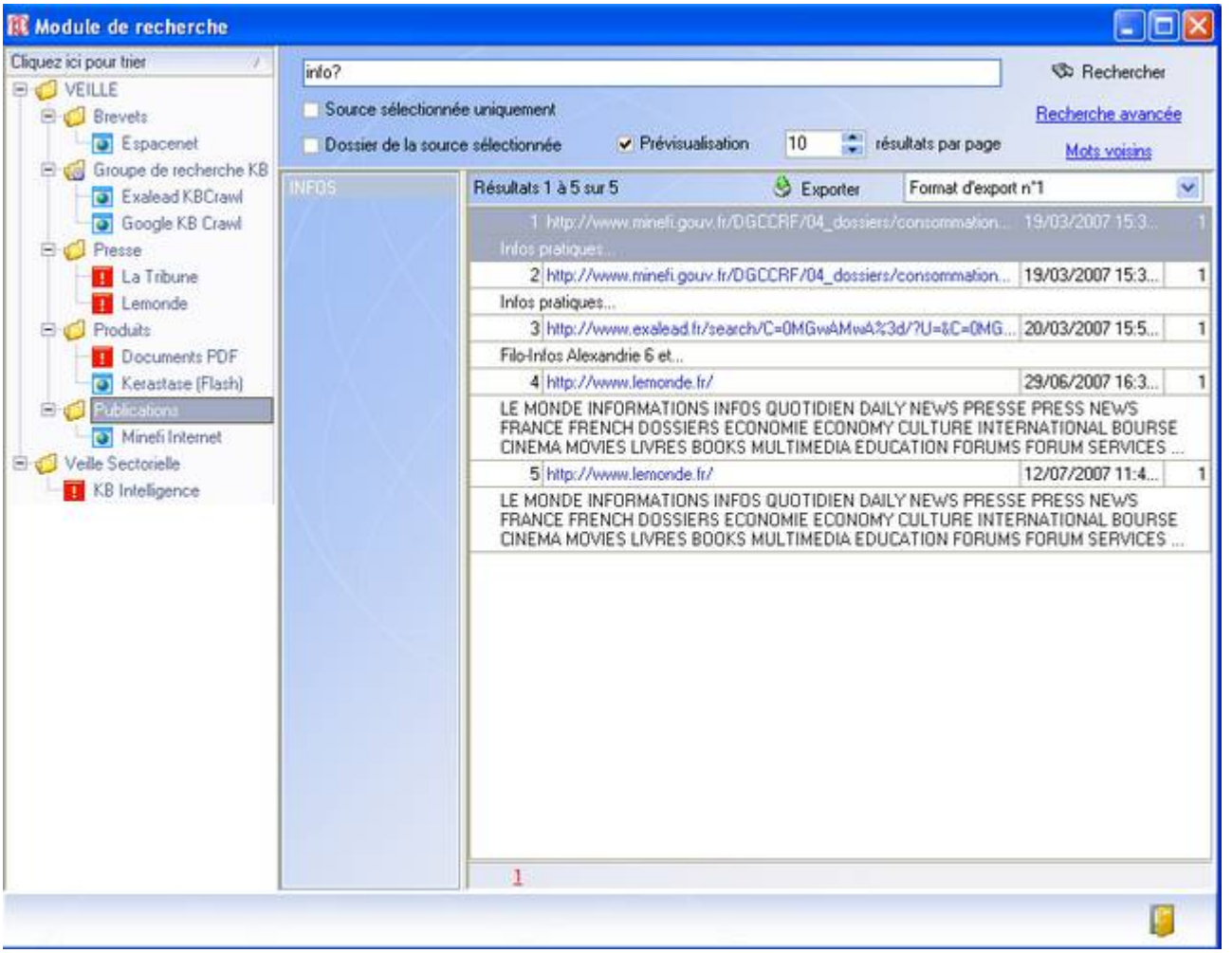


Figure 71 : Masque (exemple 1).

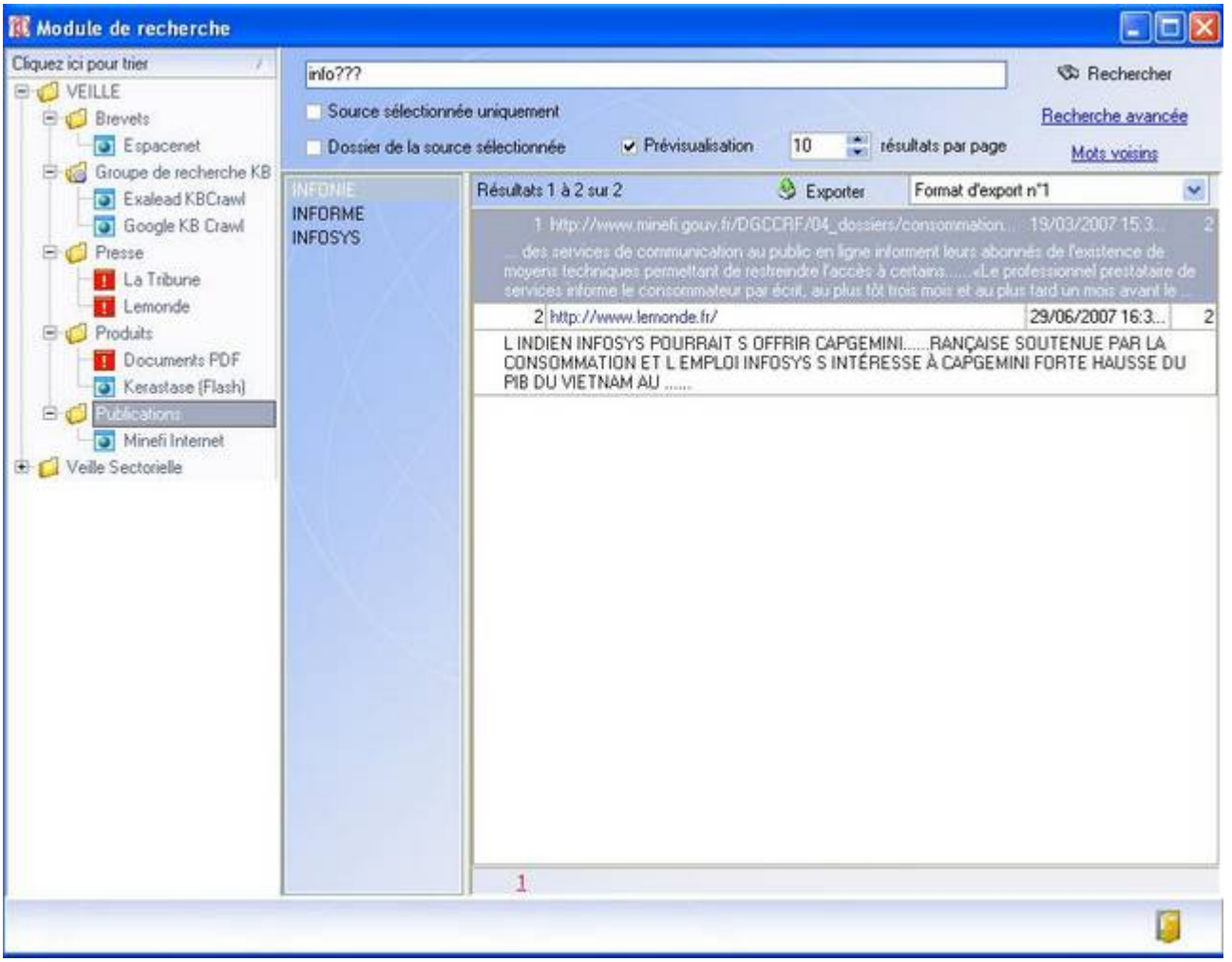


Figure 72 : Masque (exemple 2).

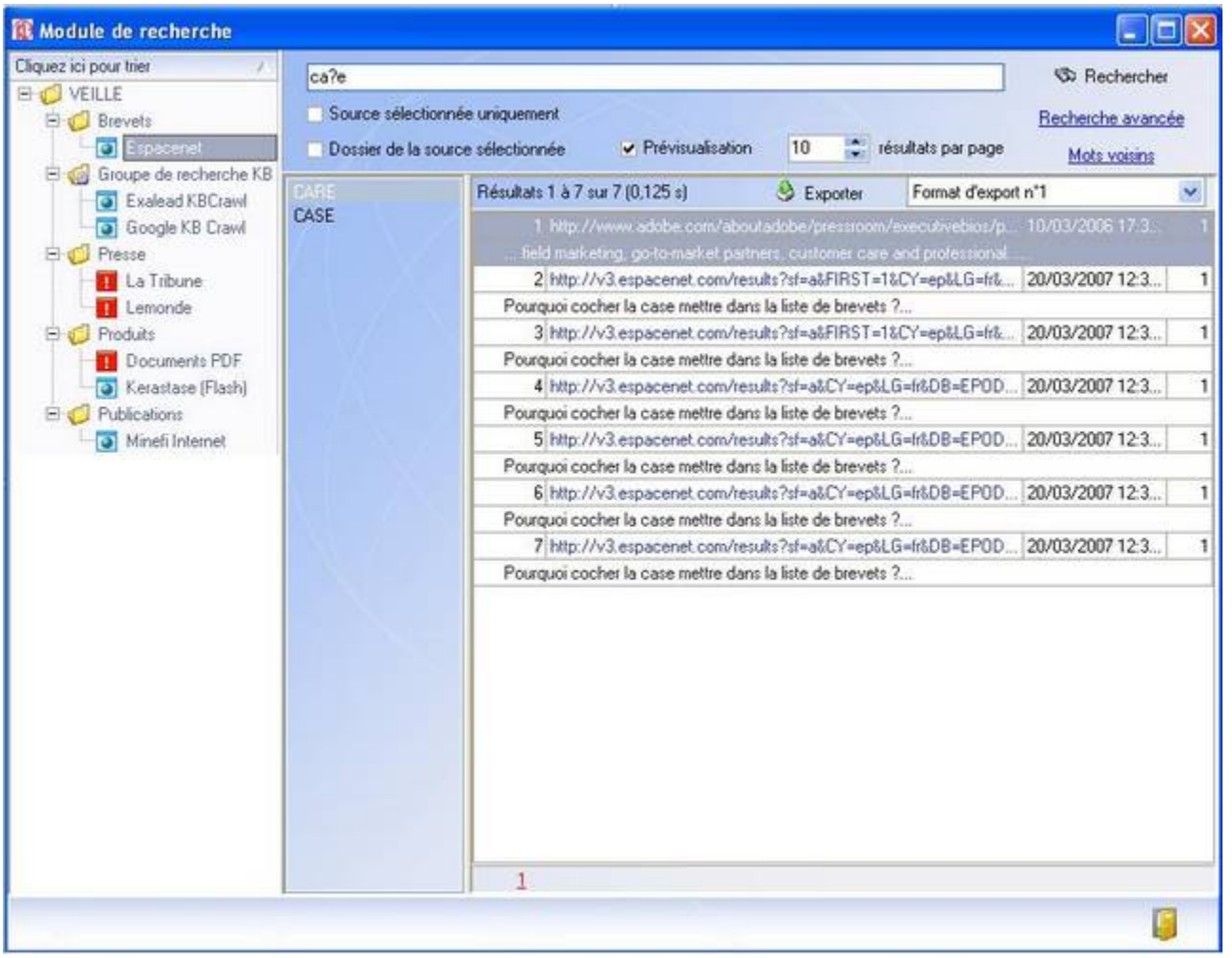


Figure 73 : Masque (exemple 3).

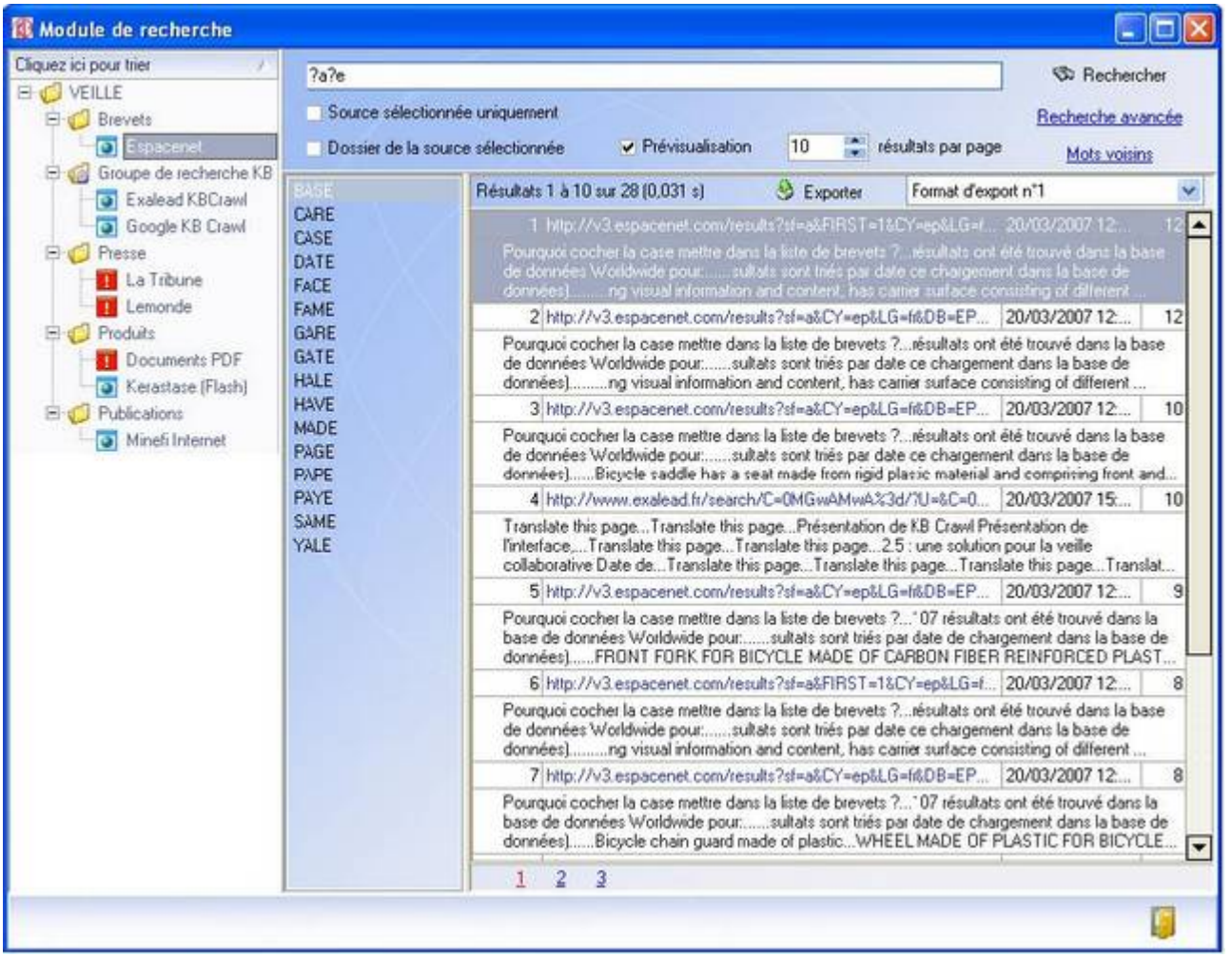


Figure 74 : Masque (exemple 4).

10.4 L'assistant de recherches avancées

Dans le cadre du haut du moteur de recherche, le bouton « Recherche avancée » permet d'accéder à une fiche qui permet de créer facilement des requêtes avancées sans avoir à se soucier des parenthèses et de la syntaxe en général.

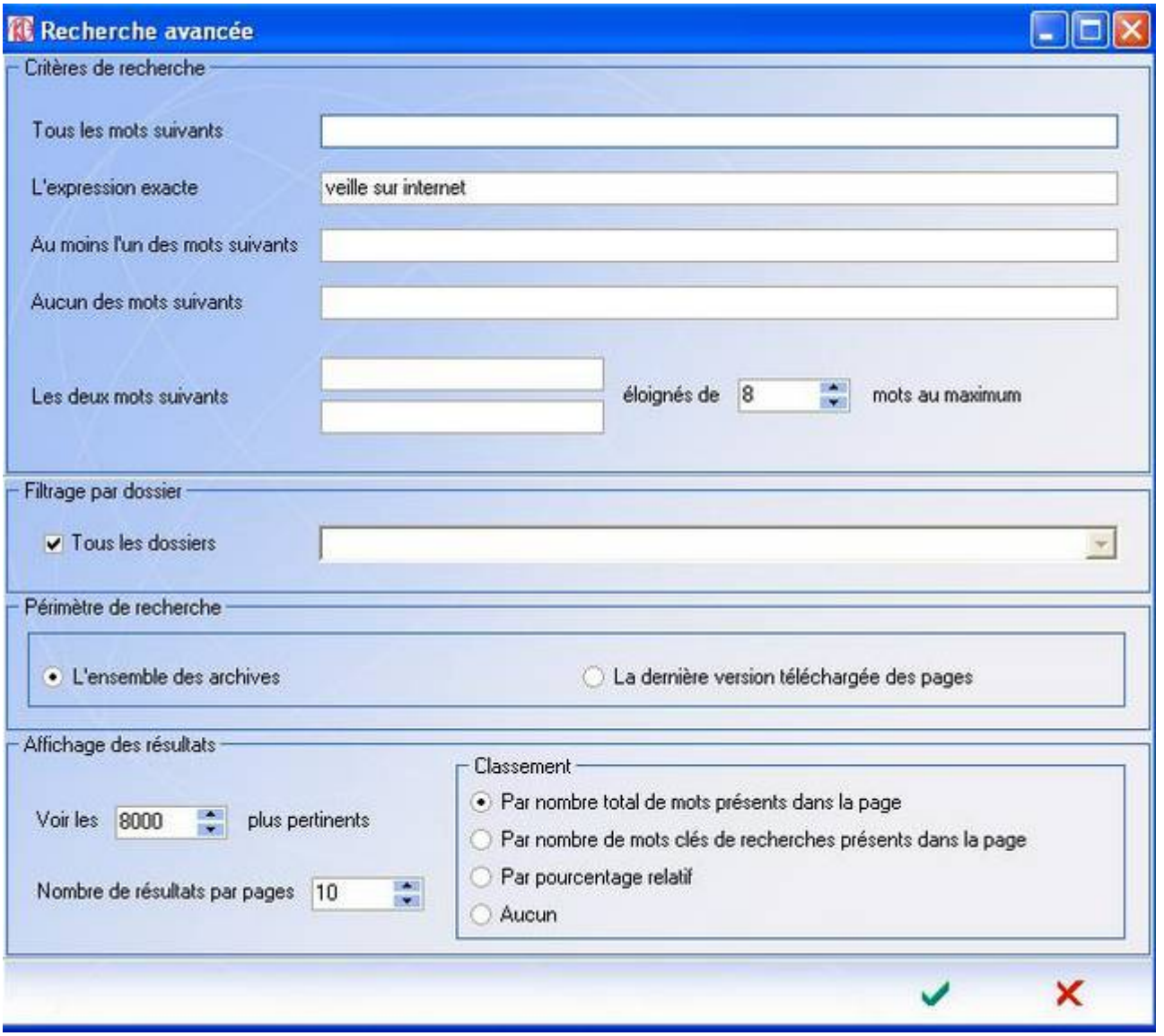


Figure 75 : Assistant à la création de requêtes avancées.

10.4.1 Le constructeur d'expressions booléennes

Les quatre premières zones de saisie rappellent l'ergonomie des formulaires de saisie de recherche avancée des moteurs de recherche de type Google, Yahoo, etc.

Une fois les expressions et mots-clés saisis dans ces zones et après avoir cliqué sur le bouton « valider », l'assistant construit automatiquement la requête qui convient au moteur de recherche.

Un peu plus bas, on trouve trois zones de saisie qui servent à faire une recherche de proximité. Il s'agit de rechercher des documents qui contiennent deux mots proches de « n » mots.

Exemple :

Lorsque l'on valide, l'expression suivante est automatiquement générée :

L'opérateur NEAR sépare deux termes qui font l'objet de la proximité.

Remarque : Des parenthèses sont automatiquement ajoutées autour de l'expression générée.

10.4.2 Filtrage par dossiers

On peut restreindre la recherche à un dossier en particulier.
Pour cela, il suffit de décocher la case « Tous les dossiers » et de sélectionner le dossier qui fait l'objet du filtrage.
Le filtre est désactivé dès lors que le bouton « Rechercher » est de nouveau sollicité.

10.4.3 Périmètre de recherche

Les pages indexées par le moteur de recherche sont toutes stockées dans la table d'archive de la base de données. Par défaut, la recherche s'effectue sur l'ensemble des pages stockées dans cet espace, sans tenir compte du type d'archivage (page de référence, dernière version, versions intermédiaires).
Il est toutefois possible de restreindre le périmètre de recherche à la dernière version archivée des pages en cochant le bouton radio « La dernière version téléchargée des pages ».

10.4.4 Affichage des résultats

Par défaut, le moteur de recherche renvoie toutes les pages qui correspondent à la requête de recherche. On peut par contre restreindre l'ensemble de résultats affichés aux « n » plus pertinents.

En effet, les résultats sont par défaut classés afin d'obtenir en première page ceux qui correspondent le plus à la recherche.

Ce classement peut se faire selon plusieurs critères :

- Par nombre total de mots présents dans la page
  Ce nombre s'obtient en additionnant le nombre d'occurrences dans la page de chacun des mots-clés.

- Par nombre de mots-clés de recherche présents dans la page
  Ce nombre correspond au nombre de mots-clés de recherche différents trouvés dans la page (pertinent dans le cas de requêtes contenant l'opérateur « OR »).

- Par pourcentage relatif
  Le résultat qui contient le plus grand nombre de mots présents dans la page est classé en premier et possède 100%. Ensuite, les autres résultats possèdent un pourcentage relatif calculé en fonction du nombre de mots-clés de recherches présents dans la page comparativement au premier du classement.

On peut également choisir de n'avoir aucun classement, ce qui optimise le temps d'exécution de la recherche car le moteur de recherche n'a pas à compter les mots dans ce cas.

La dernière option d'affichage concerne le nombre de résultats affichés par page.

11 Le journal

Depuis le menu affichage, il est possible de consulter le journal dans lequel s'inscrivent principalement des informations relatives aux crawls réalisés :

- un rapport pour chaque page crawlée,
- une notification des liens ignorés,
- un récapitulatif en bas de page qui rapporte :
    - le total de pages crawlées avec succès,
    - le total de liens ignorés,
    - le total de pages non trouvées,
    - le total pour d'éventuelles autres anomalies (échec de parsing, échec au moment du stockage dans la base…)
- un rapport d'éventuelles anomalies évoquées plus haut.

Exemple classique de journal : après le crawl du site de TF1.

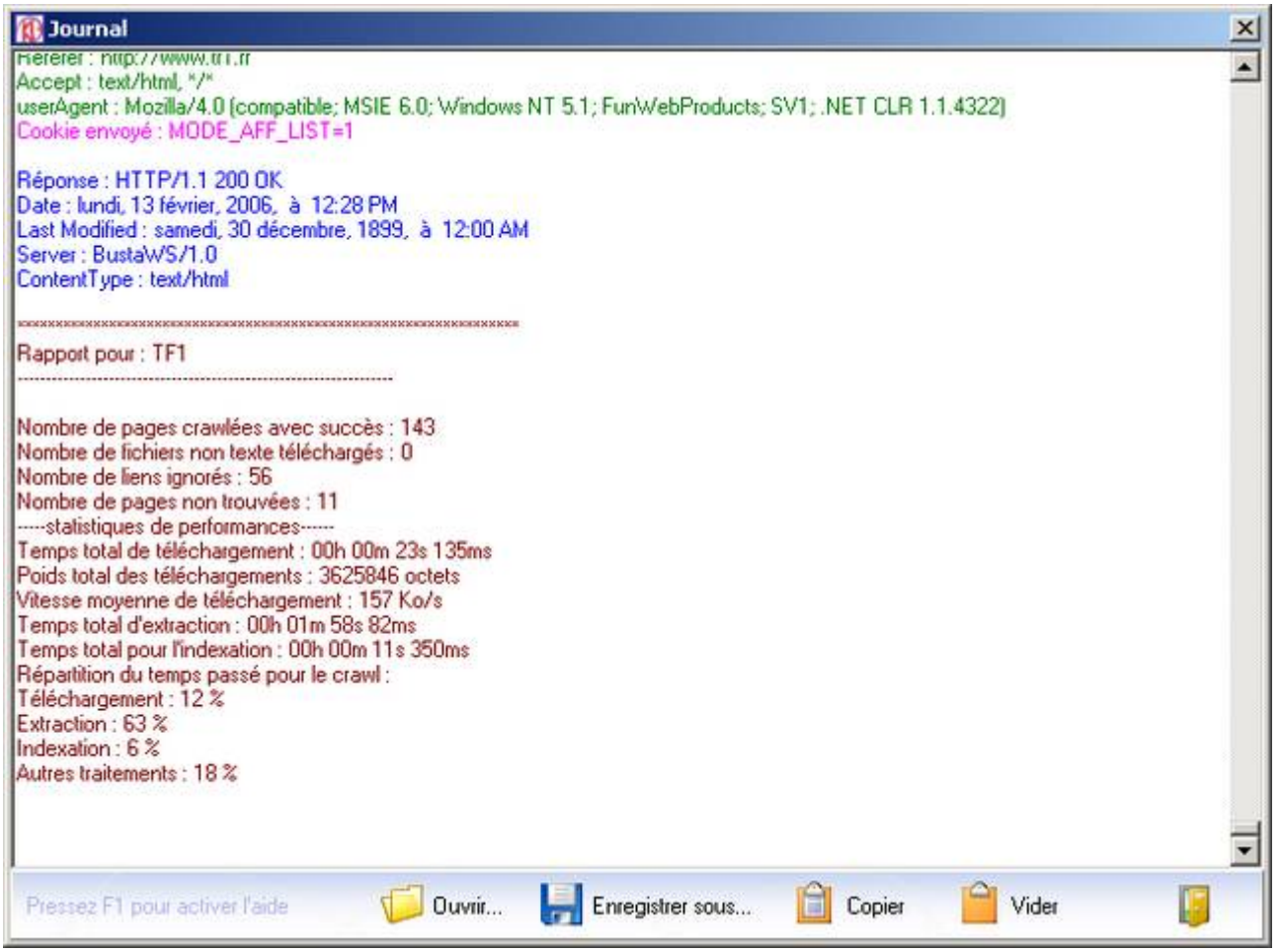


Figure 76 : Visualisation du journal d'un crawl.

En synthèse, on voit qu'il y a des pages non trouvées et beaucoup de liens ignorés.

Analyse du journal :

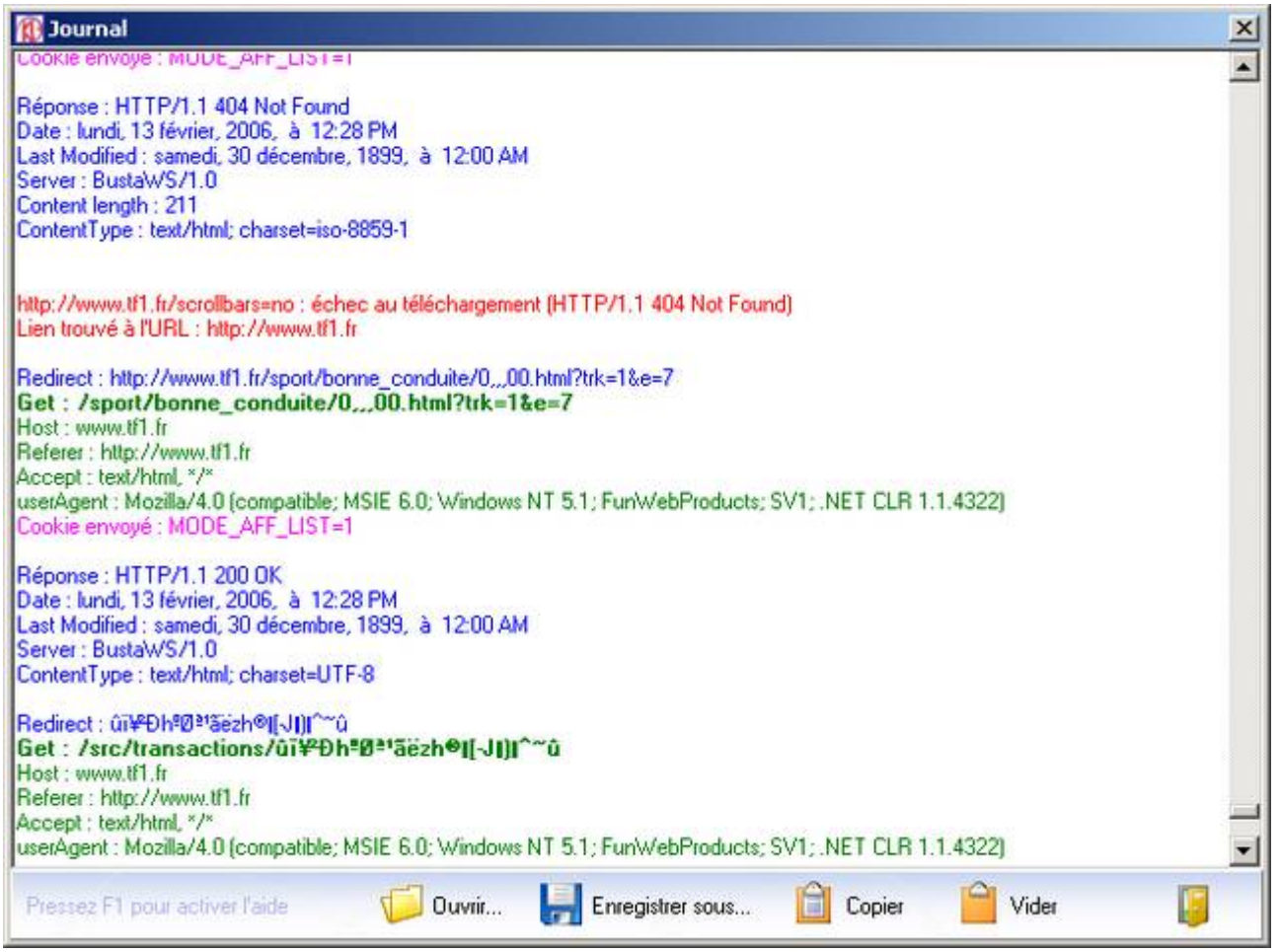


Figure 77 : Journal (page non trouvée).

Il y a des pages non trouvées :

http://www.TF1.fr/scollbarw=no : échec au téléchargement (HTTP/1.1 404 Not Found)
Lien trouvé à l'URL : http://www.TF1.fr

L'extracteur de liens de KB Crawl a interprété « scollbars=no» comme étant un lien. C'est une chaîne extraite d'un script JavaScript.
Cette chaîne était passée en paramètre d'une fonction Javascript navigante et le parser n'a aucun moyen de savoir lequel des paramètres de cette fonction est l'URL, il prend donc tous les paramètres et les collectionne en tant que liens.
Ce n'est qu'au téléchargement que l'on s'aperçoit que ce lien n'est pas valide : la page n'a pas été trouvée. Cet échec est inscrit au journal mais ne constitue pas une anomalie.
Comme le dernier tri des liens valides et non valides se fait au moment du téléchargement (1.4), cette ligne du journal est très fréquente.
Il peut aussi s'agir d'un lien mort : Le document vers lequel mène l'URL n'est pas disponible (supprimée, déplacée, etc.).
Connaître la page mère d'un lien « mort » peut être précieux pour le gestionnaire d'un site Web par exemple qui détecte dans ce cas une anomalie à l'intérieur du site qu'il maintient.

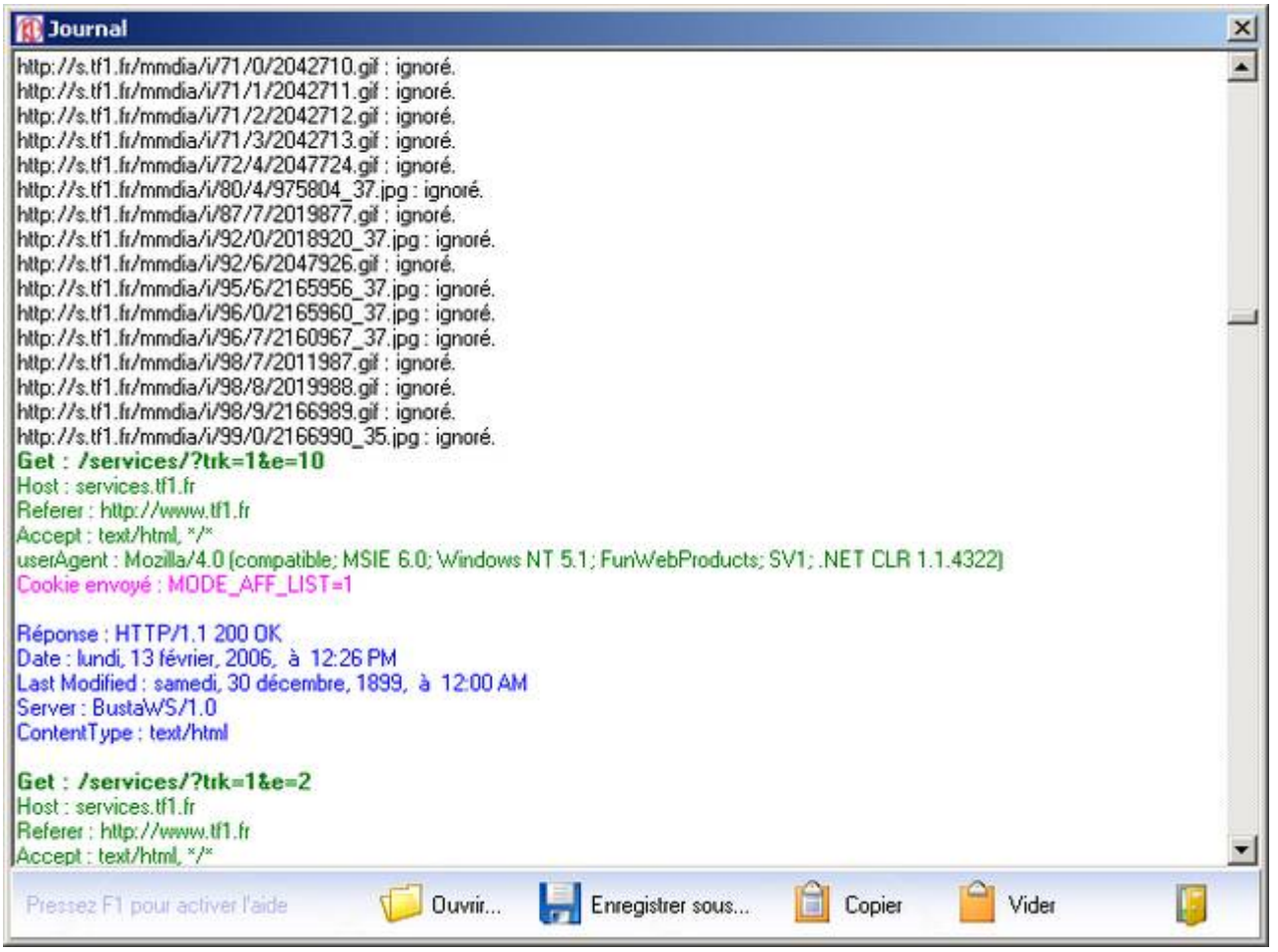


Figure 78 : Journal (fichier ignoré).

http://s.tf1.fr/mmdia/i/98/5/729985.gif : ignoré.

L'extension « gif » ne fait pas partie de la bibliothèque d'extensions de la source TF1, le lien est donc ignoré (une image à l'intérieur d'une page Web a toujours un lien).

Certains sites appellent des adresses aux extensions qui leur sont spécifiques. Il est donc primordial de connaître les liens ignorés pour éventuellement ajouter des extensions spécifiques à la bibliothèque d'extensions.

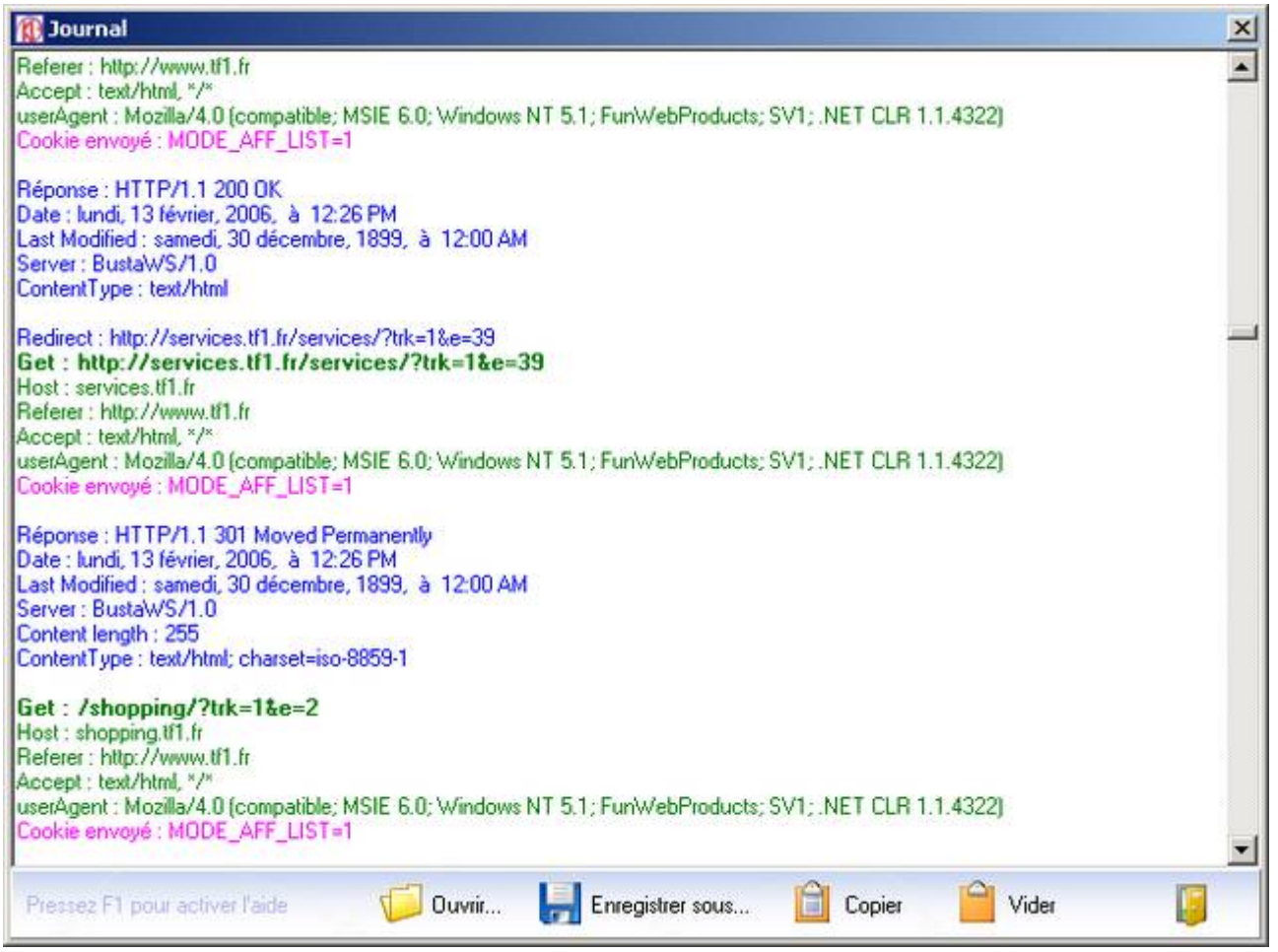


Figure 79 : Journal (téléchargement avec succès).

Quelques exemples d'URL crawlées avec des redirections.

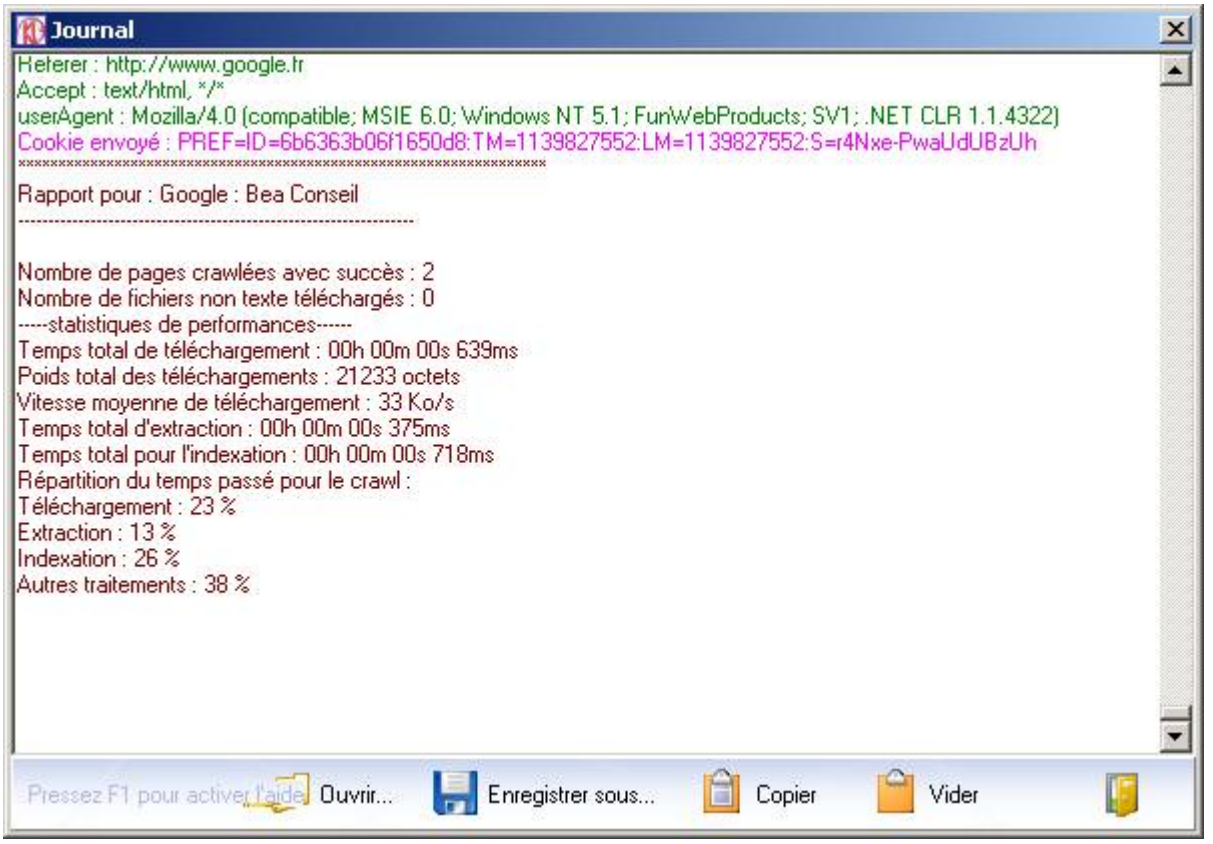


Figure 80 : Journal (compte-rendu).

Voici un compte rendu de crawl sur le site Google :

Les dernières lignes inscrites au journal lors d'un crawl présentent un rapide compte-rendu du crawl.
Les premières lignes de ce compte-rendu indiquent les proportions de documents téléchargés ou non.
Ensuite, un compte rendu informe des performances du crawl, notamment de la vitesse de téléchargement.
Il est important de remarquer que malgré une vitesse de téléchargement théorique de 150Ko/s, la vitesse moyenne du crawl n'est que de 33Ko/s.

Différents facteurs expliquent cet écart entre vitesse théorique de la ligne délivrée par le fournisseur d'accès Internet et la vitesse réelle pendant le crawl, dont les deux principaux sont :

- Le serveur requêté a une réactivité limitée et une bande passante sortante partagée par d'autres utilisateurs. Pour une même vitesse théorique, on observe des vitesses moyennes de téléchargement tout à fait différentes selon les serveurs.

- La bande passante délivrée par le fournisseur d'accès est partagée par d'autres utilisateurs en train de télécharger.

Le temps total d'extraction, quant à lui, est directement lié aux performances de l'ordinateur sur lequel est installé KB Crawl et qui effectue les traitements.
Il n'est pas nécessaire d'analyser le journal systématiquement, mais cela peut être utile lorsque l'on n'obtient pas immédiatement le résultat escompté et que l'on veut comprendre pourquoi, afin d'ajuster sa stratégie de crawl.

12 Options

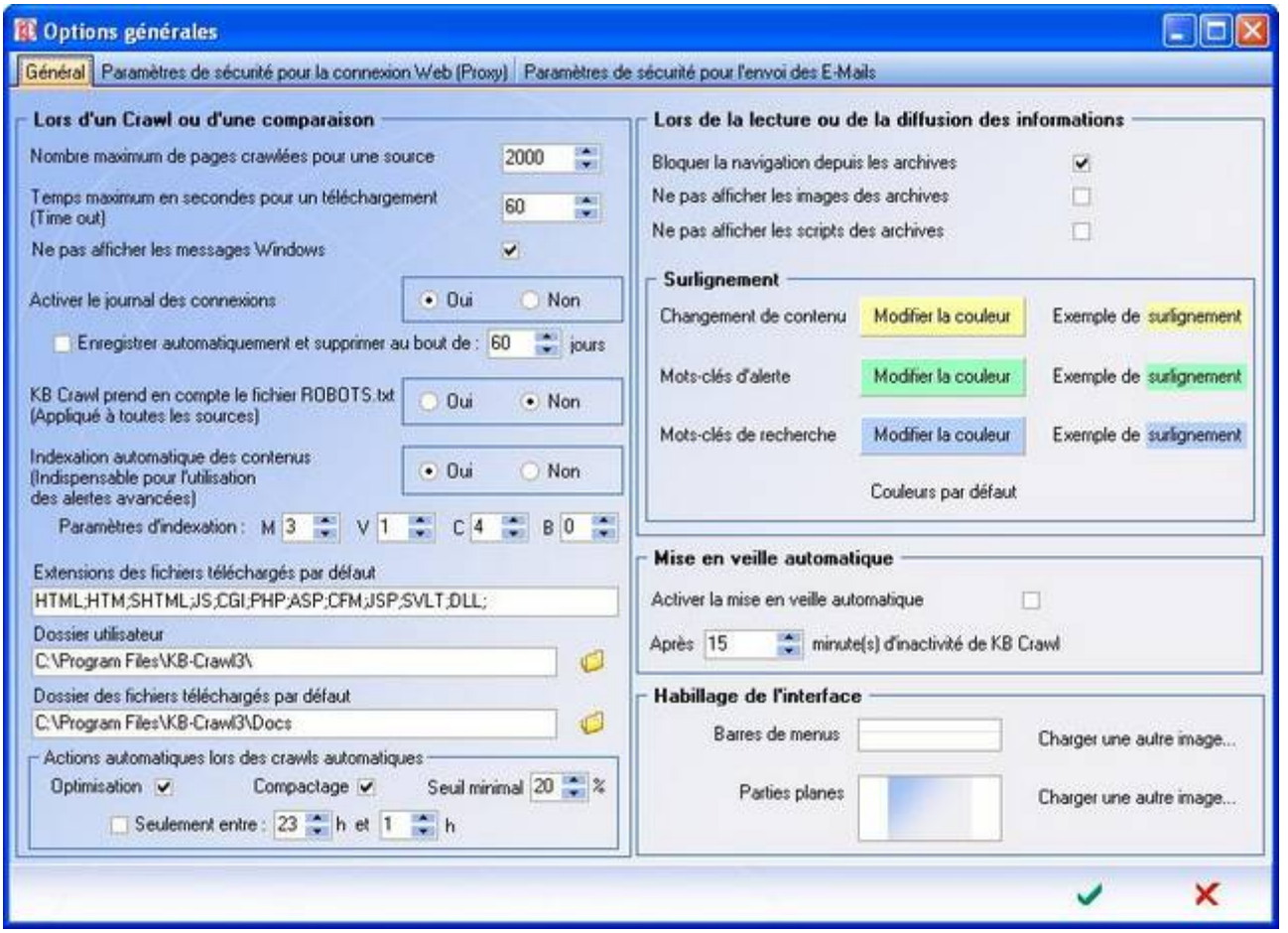


Figure 81 : Onglet "Général" du menu Options.

Le menu « Options générales » est accessible depuis la barre d'outils générale en cliquant sur le bouton « Options » 

Le menu « Options » est une fenêtre contenant trois onglets.

12.1 Onglet « général »

- Nombre maximum de pages crawlées pour une source
  Indiquer ici la limite pour le nombre total de pages explorées lors d'un crawl quelles que soient les valeurs indiquées pour les profondeurs de page et de site.

- Temps maximum en secondes pour un téléchargement (Time out)
  En anglais, « Time Out » : temps limite pour une tentative de téléchargement de page ou fichier. 30 secondes est la valeur conseillée.

- Ne pas afficher les messages Windows
  Mode silencieux : permet ou non que des boites de dialogues (authentifications, messages d'erreurs divers) surgissent des navigateurs intégrés à KB Crawl.
  Habituellement, il est conseillé d'utiliser le mode silencieux. Si la connexion Internet de l'ordinateur requiert une authentification systématique pour l'accès à un serveur proxy notamment, il peut s'avérer nécessaire de désactiver le mode silencieux.

- Activer le journal des connexions
  Permet d'activer ou de désactiver le journal(11).
  Cette fonctionnalité a été ajoutée dès la version KB Crawl 2.1 pour assurer la compatibilité avec Windows 98 (dans ce cas il faut désactiver le journal).

- Enregistrer automatiquement et supprimer au bout de X jours
  Permet d'automatiser l'enregistrement du fichier journal et d'en conserver l'historique sur une durée choisie.
  L'enregistrement se fait automatiquement lors du basculement en mode de crawl automatique, lors de l'arrêt de KB Crawl et à minuit si KB Crawl est en mode automatique.
  Les fichiers journaux sont stockés dans le dossier Utilisateur et portent l'extension RTF.

- KB Crawl prend en compte le fichier ROBOTS.txt (Appliqué à toutes les sources).
  Lorsque la case « Prendre en compte pour toutes les sources » est cochée, toutes les sources prennent en compte le fichier ROBOTS.TXT éventuellement situé à la racine du site crawlé.

- Indexation automatique des contenus (indispensable pour l'utilisation des alertes avancées)
  Permet l'indexation systématique des contenus après chaque crawl. Cette option est indispensable pour l'utilisation des alertes avancées.

- Paramètres d'indexation
  Il est possible de modifier les paramètres du moteur d'indexation de KB Crawl grâce à ces 4 champs. Le champ M correspond au nombre minimum total de caractères dans un mot; le champ V correspond au nombre minimum de voyelles dans les mots-clés ; le champ C correspond au nombre minimum de caractères à partir duquel le contrôle du nombre de voyelles est réalisé ; enfin, le champ B correspond à l'écart entre le nombre de caractères numériques et alphabétiques (ex : si le champ B=2, cela veut dire qu'il devra y avoir au moins 2 caractères alphabétiques de plus que de numériques ; si le champ B=-3, cela voudra dire qu'il pourra y avoir dans le mot jusqu'à 3 caractères numériques de plus que d'alphabétiques.)

- Extensions des fichiers téléchargés par défaut
  Cette liste recense la liste des extensions correspondant aux fichiers dont on veut s'assurer qu'ils seront téléchargés et non ignorés.

- Dossier utilisateur
  Dossier contenant tous les fichiers accessible en lecture et écriture par l'utilisateur de KB Crawl (manuel, fichiers journaux, fichiers XSLT, etc.).
  Par défaut, il s'agit du dossier d'installation de KB Crawl, mais il peut être utile de le modifier si l'utilisateur a des droits limités sur son poste.

- Dossier des fichiers téléchargés par défaut
  Ce champ permet de choisir le dossier dans lequel l'on souhaite enregistrer les fichiers téléchargés par KB Crawl (.doc, .ppt, etc.).

- Actions automatiques lors des crawls automatiques
  Lorsque KB Crawl fonctionne en mode automatique, il est possible d'optimiser la base d'archives et/ou de faire un compactage de la base de données automatiquement, si la taille de celle-ci augmente trop vite. Pour cela, il est nécessaire de cocher les cases correspondantes à son choix, et de choisir un seuil à partir duquel ces actions se réaliseront automatiquement.

- Seulement entre X h et X h
  En cochant cette case, il est possible de choisir une fenêtre temporelle pendant laquelle les actions automatiques (optimisation et/ou compactage) pourront s'enclencher.

- Bloquer la navigation depuis les archives
  Lorsque l'on est dans l'explorateur d'archives, il est possible de bloquer la navigation depuis la page que l'on visualise dans le navigateur intégré à KB Crawl :
  - les hyperliens du document visualisé deviennent inactifs
  - le document visualisé est le reflet strict de la source stockée dans la base de données : les frames par exemple, ne sont pas téléchargés.

- Ne pas afficher les images des archives
  Lorsque l'on est dans l'explorateur d'archives, il est possible de ne pas afficher les images, cela peut être utile, notamment lorsque la connexion à Internet est lente, pour un affichage des informations plus rapide.

- Ne pas afficher les scripts des archives
  Lorsque l'on est dans l'explorateur d'archives, il est possible de ne pas afficher les scripts qui s'exécutent sur les pages que l'on visualise, cela peut permettre une visualisation plus claire.

- Couleur de surlignement
  Couleur utilisée pour le surlignement :

  - des changements de contenu dans une page
  - des mots-clés d'alerte
  - des mots-clés de la recherche.

  Il est possible de changer ces couleurs en cliquant sur le bouton « Modifier la couleur », et de visualiser un exemple de surlignement sur la droite.

- Activer la mise en veille automatique
  Lorsqu'il n'y a plus eu aucune interaction entre l'utilisateur et KB Crawl pendant une durée (que l'on paramètre dans le champ juste au-dessous), la surveillance automatique peut se déclencher automatiquement. Pour activer ce mécanisme automatique, il suffit de cocher cette option.

- Habillage de l'interface
  Il est possible de personnaliser l'interface de KB Crawl en plaçant les images de fonds autres que celles fournies avec le logiciel, on peut affecter un « papier peint » aux barres de menus ainsi qu'aux surfaces planes que l'on retrouve sur tous les écrans.

12.2 Onglet « Paramètres de sécurité pour la connexion Web (Proxy)

Si l'accès Internet de l'ordinateur qui utilise KB Crawl passe par un serveur proxy, il est nécessaire de cocher la case «La connexion Internet utilise un serveur Proxy».
KB Crawl détecte automatiquement les paramètres du serveur Proxy utilisé si ceux-ci sont spécifiés dans les options de connexion d'Internet Explorer. Si ce n'est pas le cas, il faut alors renseigner les informations concernant le serveur Proxy.

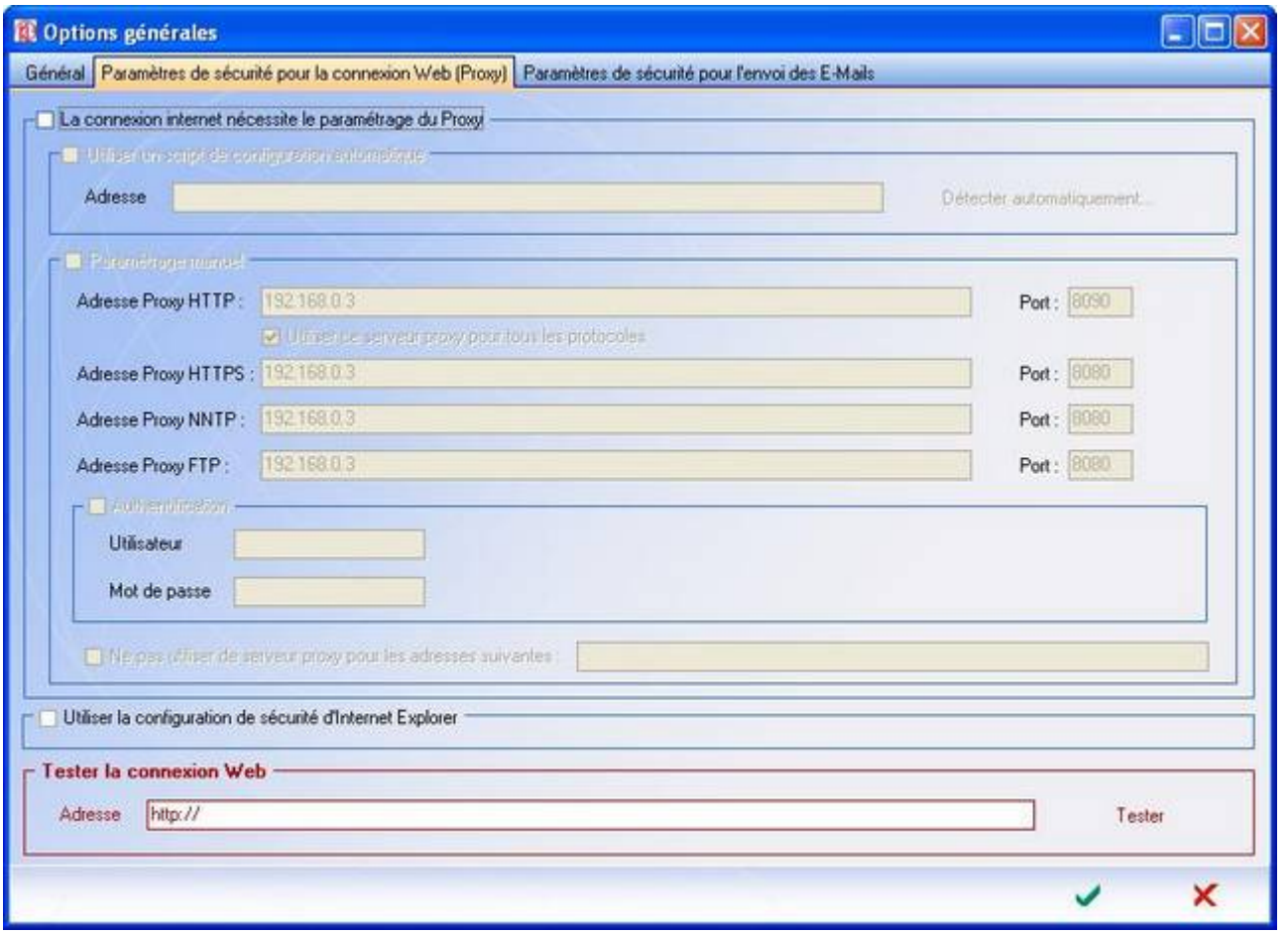


Figure 82 : Onglet "serveur proxy" du menu Options.

Deux options sont possibles :

12.2.1 Utiliser un script de configuration automatique

Saisir dans le champ « Adresse » l'adresse du script de configuration automatique.
Le bouton « Détecter automatiquement » permet de rechercher cette information dans la base de registre si elle y est enregistrée via Internet Explorer.
A chaque ouverture de session KB Crawl ou après validation des options, le script de configuration automatique est téléchargé et les paramètres du serveur proxy extraits de ce script et enregistrés.

12.2.2 Paramétrage manuel

Dans le champ « Adresse Proxy http », saisir le nom de machine du serveur ou son adresse IP.
Dans le champ « Port » saisir le port d'écoute sur ce serveur.
Si une authentification est nécessaire, cocher la case « Authentification » puis l'identifiant et le mot de passe requis pour cette authentification.

Par défaut, KB Crawl utilisera ce serveur proxy pour tous les protocoles. Cependant, il est possible d'utiliser d'autres serveurs pour des protocoles différents.
Ainsi, si l'on désire crawler des sources HTTPS, NNTP ou FTP en utilisant un serveur proxy différent que celui pour le HTTP, il est nécessaire de décocher la case « Utiliser ce serveur proxy pour tous les protocoles », et de renseigner les champs correspondants (adresse et port).

12.2.3 Utiliser la configuration d'Internet Explorer

Cette option permet de définir l'option par défaut du même nom dans toutes les sources nouvellement créées. Elle peut être utile dans certains cas très spécifiques, notamment pour autoriser KB Crawl à se connecter à des serveurs proxy comprenant des sécurités avancées.

12.2.4 Tester la connexion Web

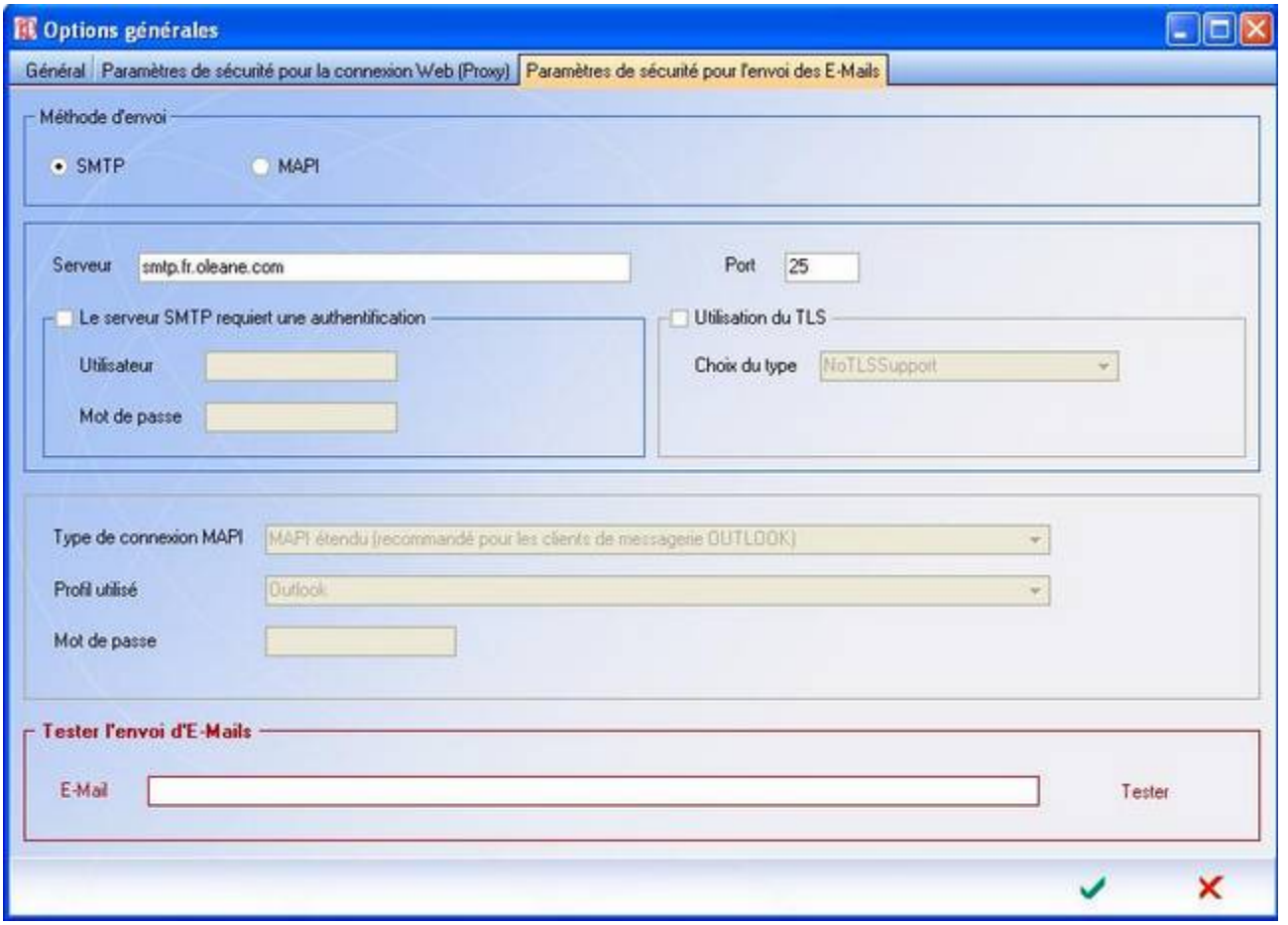
Afin de vérifier si les paramètres renseignés pour la connexion Web sont corrects, il est possible de tester la connexion. Pour ce faire, il faut renseigner le champ « Adresse » avec une URL valide, et cliquer sur « Tester ». Si les paramètres renseignés sont corrects, un message de confirmation s'affiche à l'écran (sinon, il est nécessaire de modifier les paramètres de connexion.)

12.3 Onglet « Paramètres de sécurité pour l'envoi des E-Mails

12.3.1 Envoi par connexion SMTP

Pour utiliser l'envoi d'e-mails par le protocole SMTP, cocher le bouton radio « SMTP ».

Il est ensuite nécessaire de renseigner correctement l'adresse du serveur SMTP, son port ainsi que le nom d'utilisateur et le mot de passe de messagerie requis afin que les messages puissent être envoyés.



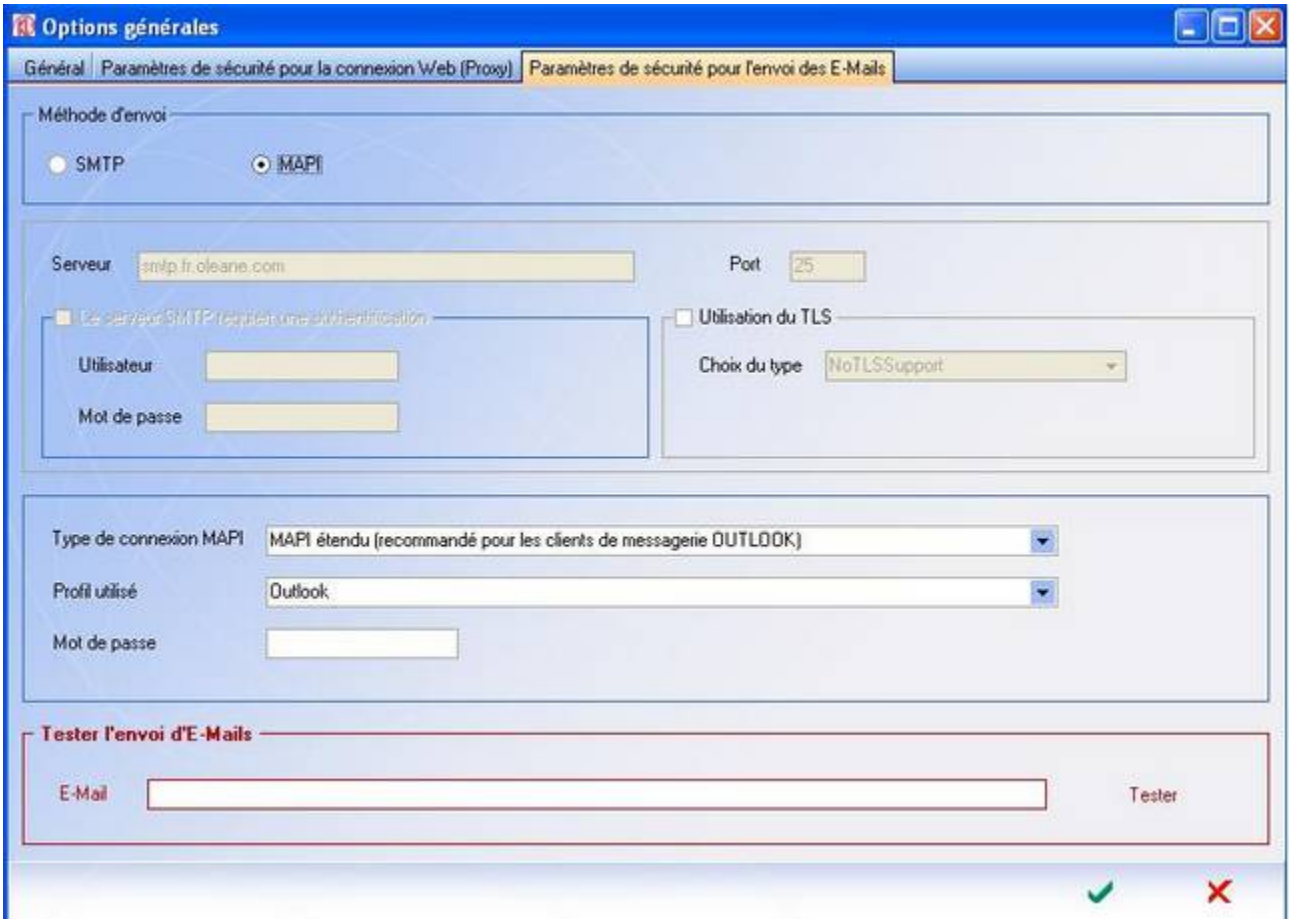
Certains serveurs de messageries, pour des raisons de sécurité, exigent une authentification il est alors nécessaire de cocher l'option « Le serveur SMTP requiert une authentification » et de renseigner les champs correspondants.

12.3.2 Envoi par connexion MAPI

Pour utiliser l'envoi d'e-mails par le protocole MAPI, cocher le bouton radio « MAPI ».

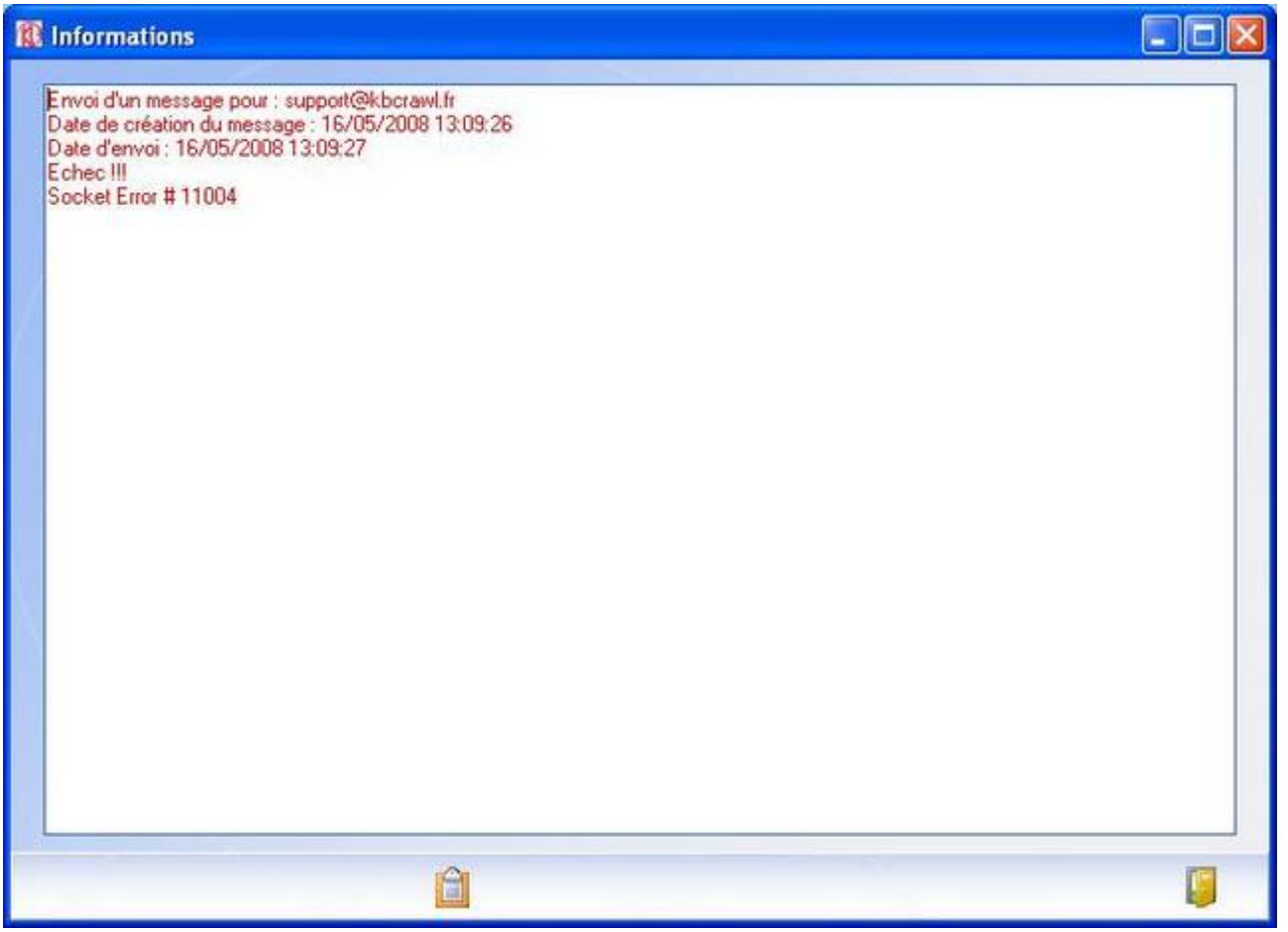
MAPI est une librairie d'applications qui communiquent avec le client de messagerie défini par défaut sur l'ordinateur : les messages sont insérés dans la boite d'envoi du client de messagerie et l'envoi de ces messages est délégué à celui-ci.

Pour OUTLOOK par exemple, il faut utiliser la librairie MAPI étendu.
Si nécessaire, choisir le profil utilisé et saisir le mot de passe de session Windows.



Afin de vérifier que les paramètres d'envoi des e-mails sont corrects, saisir une adresse e-mail valide dans la zone « E-Mail » puis cliquer sur le bouton « Tester »

En cas d'échec, le message suivant apparaît :



12.3.3 Utilisation du TLS

Il est possible d'établir une connexion en utilisant du TLS. Pour cela, il est nécessaire de cocher la case « Utilisation du TLS » et de choisir le type que l'on souhaite utiliser.

13 Fonctions utilitaires

En supplément de celles décrites dans les chapitres précédents, KB Crawl propose des fonctionnalités utilitaires accessibles depuis le menu textuel.

Voici la liste complète de ces fonctionnalités, passée en revue par élément de menu :

13.1 Fichier

13.1.1 Réduire KB Crawl en mode automatique

On peut à tout moment « fermer » KB Crawl sans que l'activité de celui-ci n'en soit perturbée.
En mode « automatique » notamment, il n'est pas utile que KB Crawl reste ouvert au même titre que les autres fenêtres d'applications.
En appelant cette fonctionnalité, la fenêtre principale de KB Crawl se ferme mais l'application « travaille » toujours et peut être rappelée à tout moment par un double clic sur l'icône située à droite dans la barre de tâches de Windows :

Si une alerte se déclenche, l'icône de KB Crawl change de couleur pour être alerté visuellement sans avoir à rouvrir la fenêtre principale de KB Crawl :

13.1.2 Quitter KB Crawl

Ferme définitivement l'application.

13.2 Edition

13.2.1 Liste des sources au format Excel

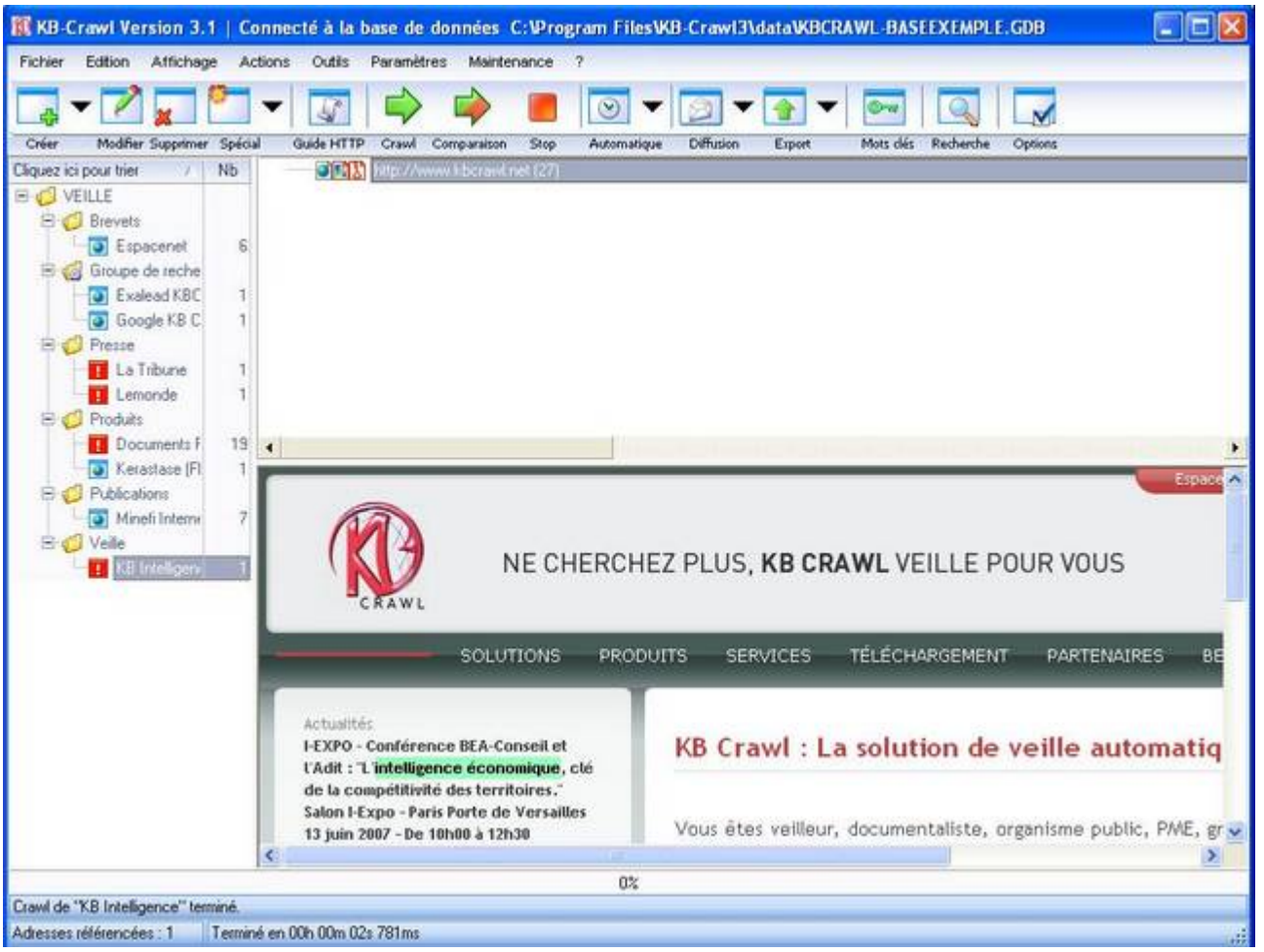
Exporte l'arborescence des sources et des dossiers au format Excel, ainsi que les URL des pages de départ :

	A	B	C	D	E	
1	Cliquez ici pour trier				Nb	URL
2	_	VEILLE			-1	
3		_	Brevets		-1	
4				Espacenet	6	http://v3.espacenet.com/results?sf=a&FIRST=1&CY=ep&LG=fr&DB=EPODOC&TI=p
5		_	Groupe de recherche KBCrawl		-1	
6				Exalead KBCrawl	1	http://www.exalead.fr/search/C=0MGwAMwA%3d/?U=&C=0MGwAMwA%3D&2q=&
7				Google KB Crawl	1	http://www.google.fr/search?hl=fr&q=kbcrawl&btnG=Recherche+Google&meta=
8		_	Presse		-1	
9				La Tribune	1	http://www.latribune.fr/rss
10				Lemonde	1	http://www.lemonde.fr/
11		_	Produits		-1	
12				Documents PDF	19	http://www.adobe.com/aboutadobe/pressroom/executivebios/main.html
13				Kerastase (Flash)	1	http://www.kerastase.ch/img/_ch/_fr/Conso/Rituals/Fermete.swf
14		_	Publications		-1	
15				Minefi Internet	7	http://www.minefi.gouv.fr/themes/technologies_info/internet/index.htm

13.3 Affichage

13.3.1 Volet de prévisualisation

Permet de montrer ou de cacher le volet de prévisualisation (cadre du bas). Lorsqu'on clique sur une URL dans l'arbre de droite, le volet de prévisualisation se met à jour afin de visualiser le document correspondant.



13.3.2 Boite à outils URL

Permet de montrer ou de cacher la « boite à outils URL ».

Celle-ci présente les fonctionnalités suivantes (que l'on retrouve dans le menu contextuel de l'arbre) :

- Page surveillée

Sélectionner une URL dans l'arbre des URL puis cliquer sur le bouton « page surveillée » pour marquer l'URL comme étant à surveiller.

- Page non surveillée

C'est la fonctionnalité inverse de la précédente, si une URL est marquée comme n'étant pas à surveiller, KB Crawl ne cherchera pas à détecter quelque changement que ce soit

dans le document, mais il effectuera le parsing pour extraire les liens, et éventuellement, les suivre ensuite si le niveau de profondeur l'indique.
Cette fonctionnalité est très utile pour surveiller des pages en profondeur dans un site sans pour autant recevoir d'alertes sur les pages intermédiaires qui servent en réalité de « pont ».

- Filtre exclusif

Rend une URL exclusive à son niveau d'arborescence.

- Filtre Black-liste

Black-liste une URL à son niveau d'arborescence.

- Filtre avancé

Donne accès au gestionnaire de filtre avancé.

- Supprimer les filtres

Supprime tous les filtres liés à la source.

- Page exportée

Marque le document sélectionné dans l'arbre d'URL comme étant à exporter.

- Page non exportée

Fonctionnalité qui a l'effet inverse de la précédente.

- Page en ligne

Ouvre le navigateur par défaut et navigue sur l'URL correspondante.

- Nouvelle source

Voir le § 4.3.16.

13.3.3 Légende

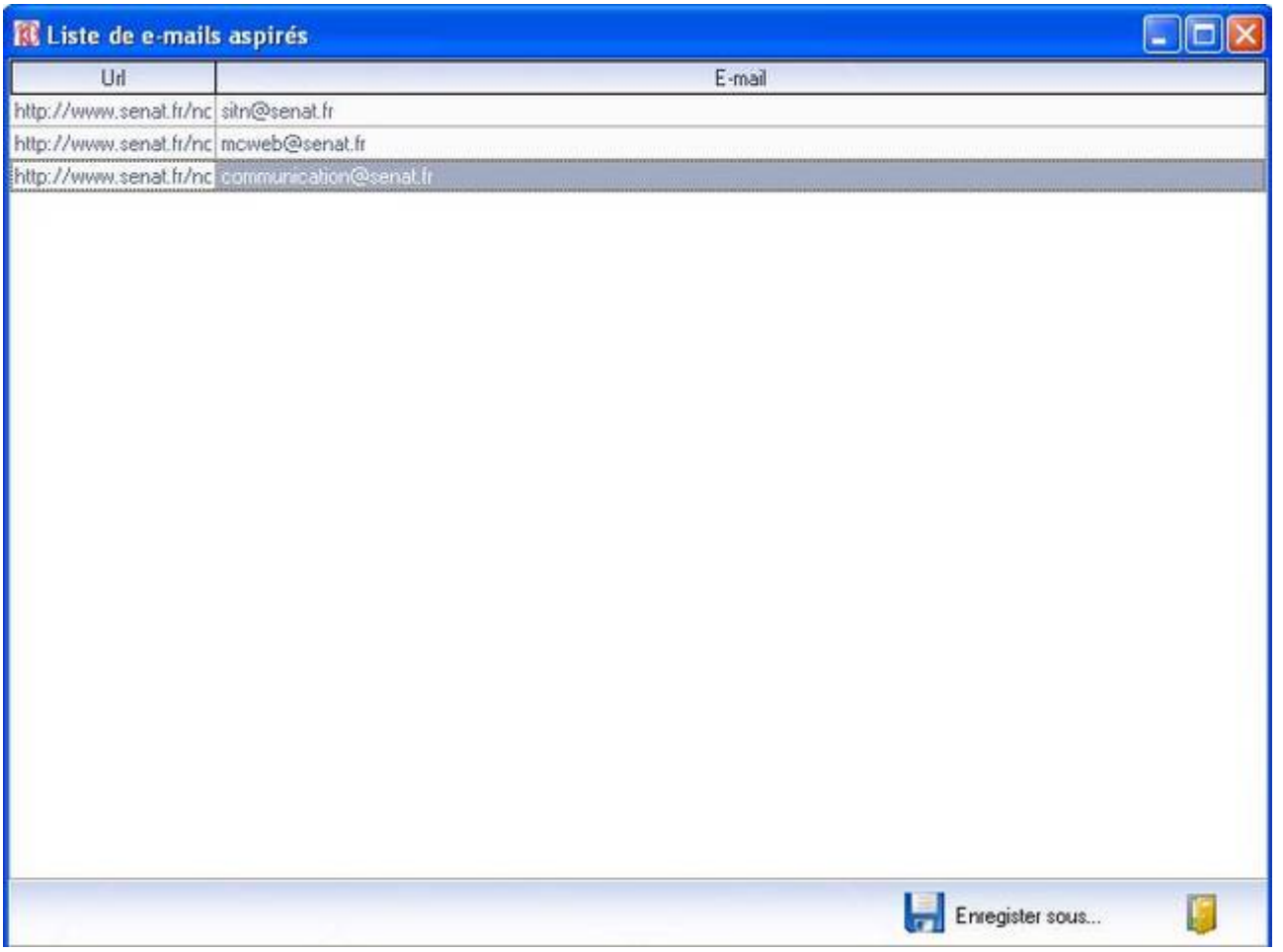
Affiche la liste des légendes descriptive des différentes icônes que l'on peut voir dans l'arbre des URL.

13.3.4 Journal

Affiche le journal des connexions (voir § 11)

13.3.5 E-mail

Montre la liste des adresses e-mail accumulées durant le dernier crawl.
Cette liste peut être exportée sous Excel.



13.3.6 KB Scraper

Ouvre la barre d'outils de KB Scraper. Pour connaître le fonctionnement de KB Scraper, se reporter à la documentation de ce module.

13.4 Actions

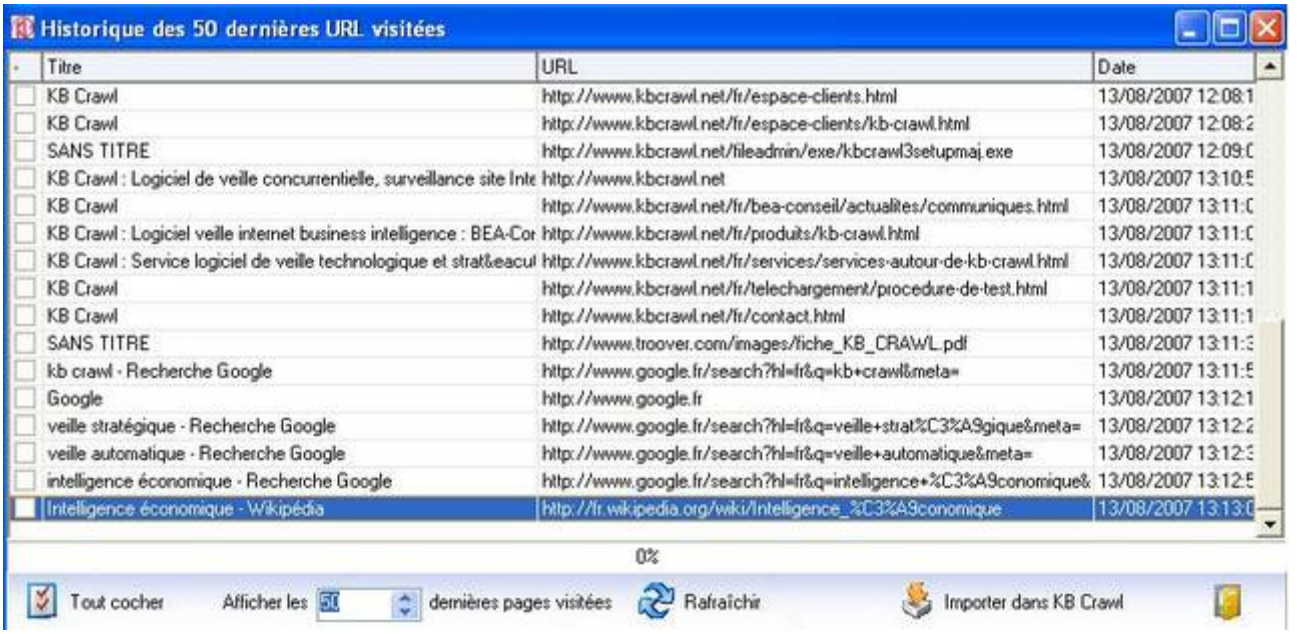
13.4.1 Installer le lien KB Crawl dans Interner Explorer

KB Crawl propose un module additionnel qui permet de récupérer l'historique des URL visitées avec Internet Explorer et de les importer sous forme de sources dans KB Crawl.

Ce module est accessible en cliquant sur un bouton dans la barre d'outils d'Internet Explorer :

En aucun cas ce bouton KB Crawl n'est installé automatiquement de manière intrusive dans Internet Explorer. Pour installer ce lien avec KB Crawl dans Internet Explorer, il faut cliquer sur l'élément de menu « Installer le lien KB Crawl ».

Lorsque l'on clique sur le bouton « KB Crawl » installé dans la barre de tâches d'Internet Explorer, cette fenêtre surgit :



Elle présente principalement une grille relatant les 50 dernières pages Web visitées avec Internet Explorer : le titre de la page, l'URL correspondante et la date de visite.

Pour sélectionner une URL, il est possible de cocher la case située à sa gauche ou d'appuyer sur la barre espace. On peut également toutes les sélectionner d'un coup en cliquant sur le bouton « Tout cocher ».

Le bouton « Rafraîchir » sert à récupérer les URL visitées après que ce module d'import ait été ouvert depuis Internet Explorer.

Lorsque les URL sont sélectionnées, cliquer sur le bouton « Importer dans KB Crawl ».
Une fenêtre surgit alors afin de spécifier le dossier de destination des sources à créer.

Cliquer ensuite sur la coche verte pour lancer l'import.

Si KB Crawl est déjà ouvert, se positionner sur le cadre de gauche où sont contenues les sources, puis appuyer sur F5 (=rafraîchir) pour faire apparaître les sources ainsi créées.

13.4.2 Déverrouiller toutes les sources

Lorsqu'un crawl est lancé pour une source, celle-ci est verrouillée jusqu'à ce que le crawl soit terminé, et ceci pour plusieurs raisons :

- empêcher qu'un autre utilisateur connecté à la même base de données ne lance un crawl sur la même source en même temps,
- empêcher qu'un autre utilisateur ne modifie les paramètres de la source pendant qu'un crawl est lancé.

Il peut arriver, si le crawl ne s'est pas terminé correctement, que la source ne soit pas déverrouillée. Dans ce cas, une icône en forme de verrou est positionnée sur la source visible depuis l'interface :

Lorsqu'une source est restée verrouillée alors qu'aucune action n'est en cours, on peut la déverrouiller très simplement en faisant un clic droit puis « déverrouiller toutes les sources ».

13.4.3 Réinitialiser les options des sources sélectionnées

Sélectionner un ensemble de sources, puis cliquer sur le bouton correspondant à cette fonctionnalité réinitialise les options des sources sélectionnées, c'est-à-dire leur donne la même valeur que lors de leur création.

13.4.4 Supprimer les archives de la source sélectionnée

Supprime toutes les archives de la source sélectionnée, sauf la version de référence de la page.

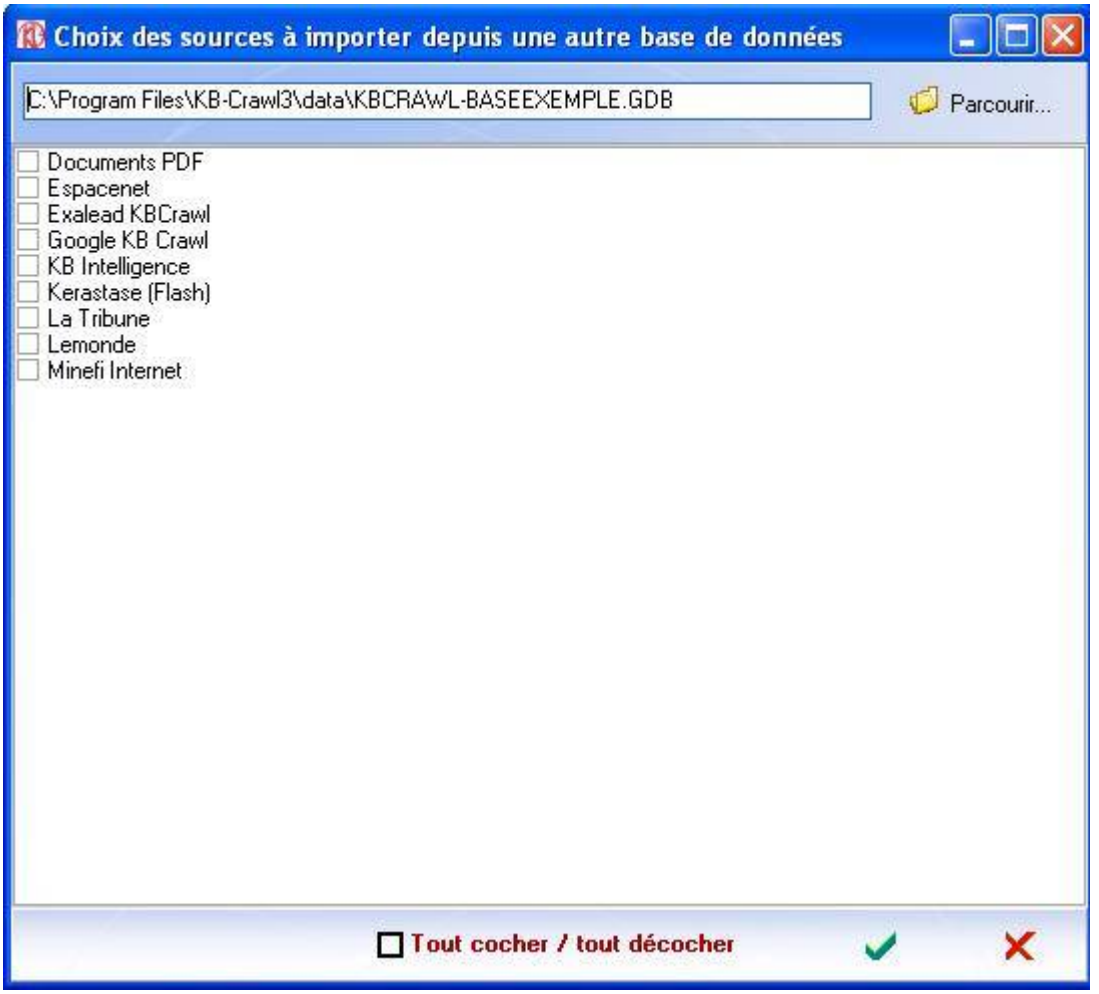
13.4.5 Initialiser toutes les connexions à la base de données

Pour réaliser certaines actions comme un compactage de la base de données par exemple, il faut que toutes les connexions clientes à la base de données soient coupées.
Il peut arriver que bien que tous les utilisateurs ne soient plus apparemment connectés à la base de données, celle-ci ait gardée en mémoire des connexions actives. Le serveur de base de données n'est pas toujours accessible, c'est pourquoi cette fonctionnalité a été créée : pour couper depuis un poste client toutes les connexions avec la base de données.

13.5 Outils

13.5.1 Importer des sources venant d'une autre base

Il est possible depuis KB Crawl d'importer des sources provenant d'une autre base de données KB Crawl. Cette fonctionnalité propose un outil simplifié pour ce type d'échange :



Dans cet écran, il suffit de choisir les sources à importer en cochant la case correspondante, puis de valider la fiche.

Les sources sélectionnées sont alors importées dans le dossier en cours dans la fiche principale de KB Crawl.

13.5.2 Importer des favoris

KB Crawl permet d'importer les favoris créés dans le navigateur Internet Explorer.
Un favori sous Internet Explorer est un fichier portant l'extension « url » placé dans un répertoire donné (le plus souvent un sous répertoire de « C:\Documents and Settings »).
Ce répertoire est enregistré dans Windows. KB Crawl le reconnaît et renvoie la liste de tous les fichiers « favoris » qui se trouvent dans ce répertoire et tous ceux de niveau inférieur.
Pour importer ces favoris, aller dans le menu «Utilitaires/Importer» de «favoris».

La fenêtre qui propose une liste de favoris à importer apparaît alors :

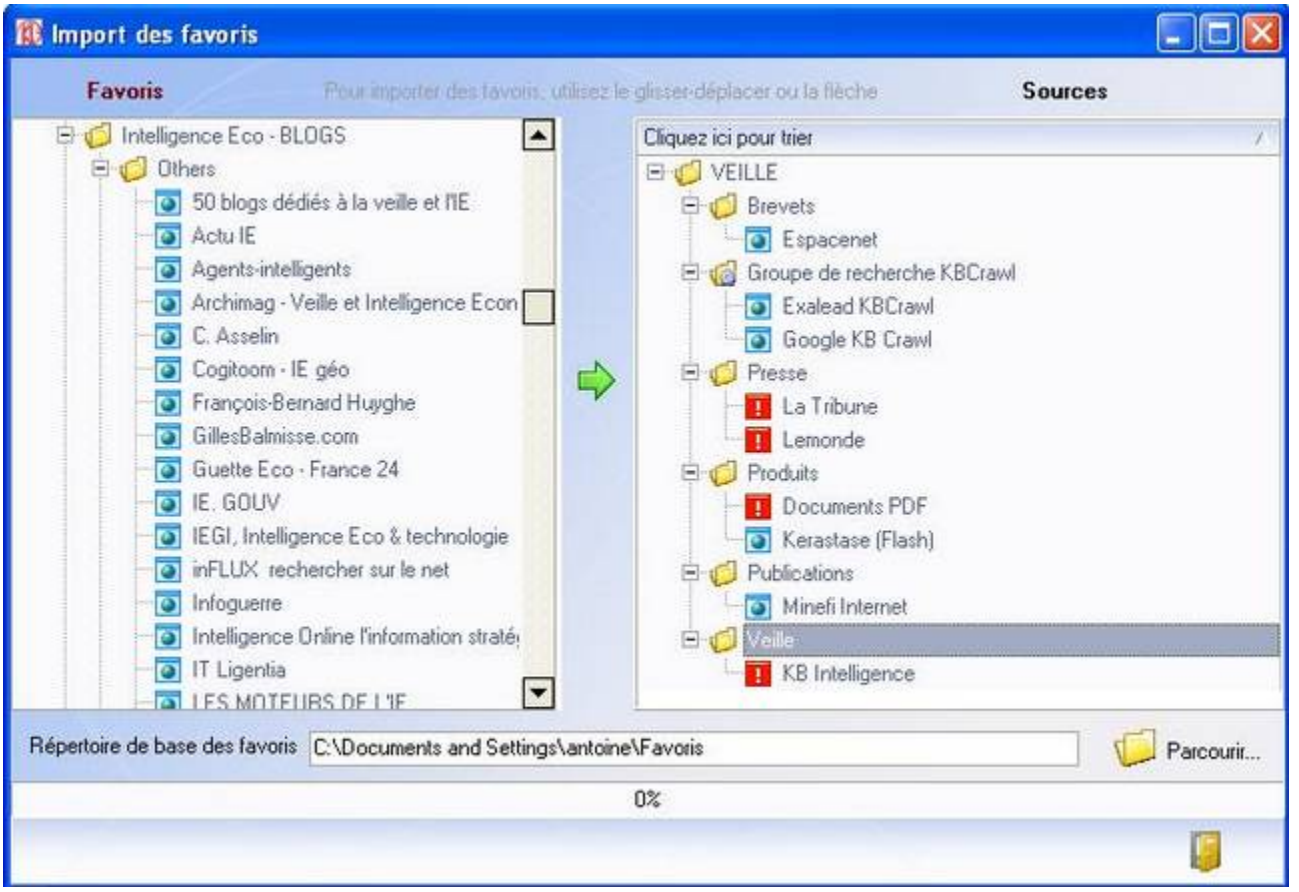


Figure 83 : Import des favoris.

Il suffit de sélectionner les favoris à importer (dans la partie gauche), de sélectionner le dossier de destination et de cliquer sur la flèche verte centrale.
Vous pouvez également utiliser le glisser/déplacer.
Les dossiers des favoris seront repris en tant que nouveaux dossiers dans KB Crawl.

13.5.3 Import-Export des sources avec KB Exchange

Le module KB Exchange est un module sophistiqué, plus complet que le précédent qui permet de gérer les échanges de données entre différentes bases de données KB Crawl.
Pour plus d'informations, voir la documentation complète du module KB Exchange.

13.5.4 Importer des sources venant d'un fichier

Il est possible d'importer des sources venant d'un fichier Microsoft Excel, CSV ou OPML.
Dans le fichier Excel, la première colonne correspond au nom de la source ; la deuxième à l'adresse URL de la page de départ. Chaque ligne correspond à la création d'une source. Dès la présence d'une ligne ne comportant pas d'adresse URL, l'import se termine.

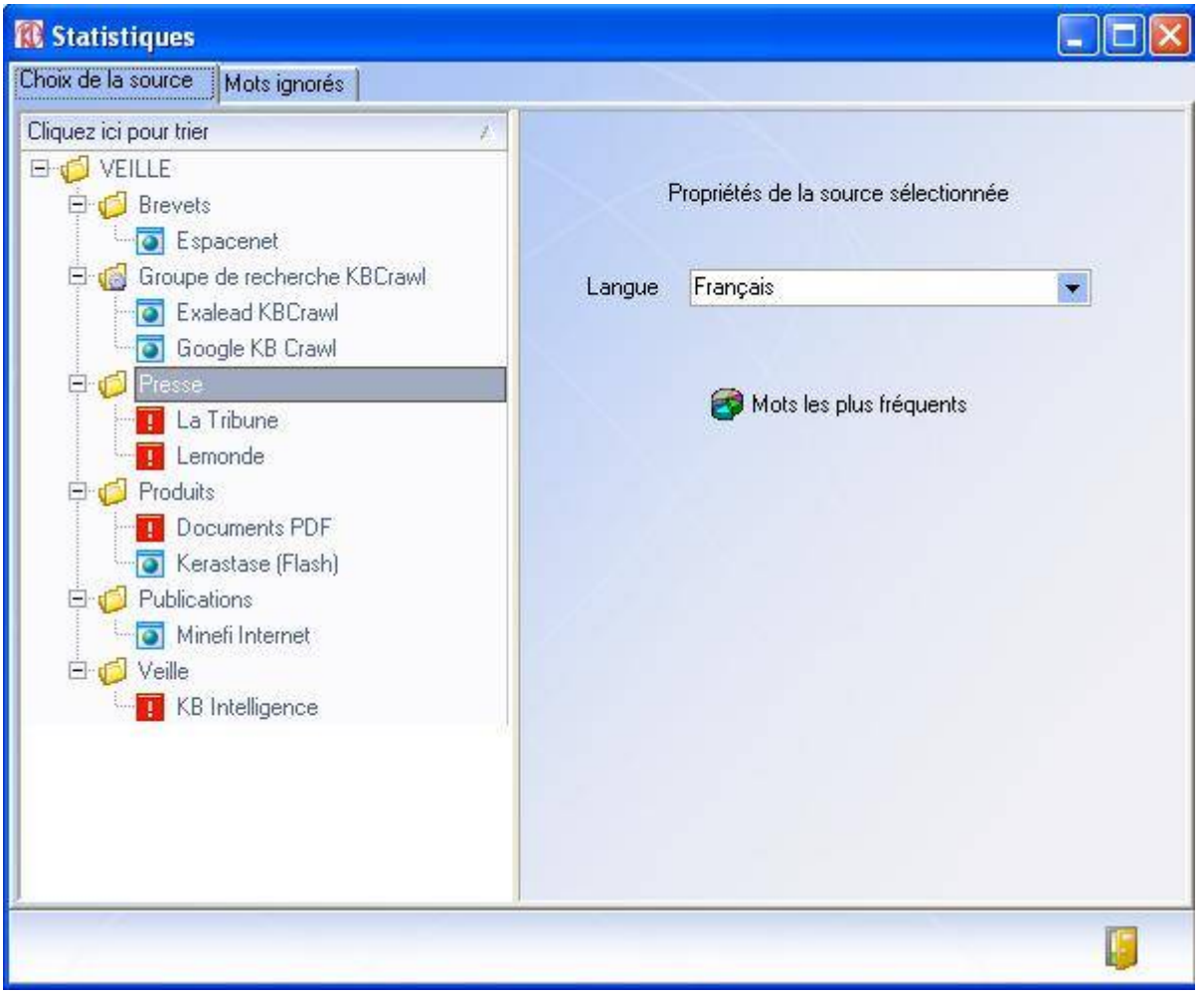
Dans le fichier CSV, chaque ligne correspond à une source à importer. Le séparateur de colonne est les points virgules. Voici un exemple de ligne :
Mon site;http://www.monsite.com

Le fichier OPML est un fichier spécifique au format XML qui contient une liste de sites, généralement des flux RSS. KB Crawl, lors de ses imports, sait gérer ce genre de formats.

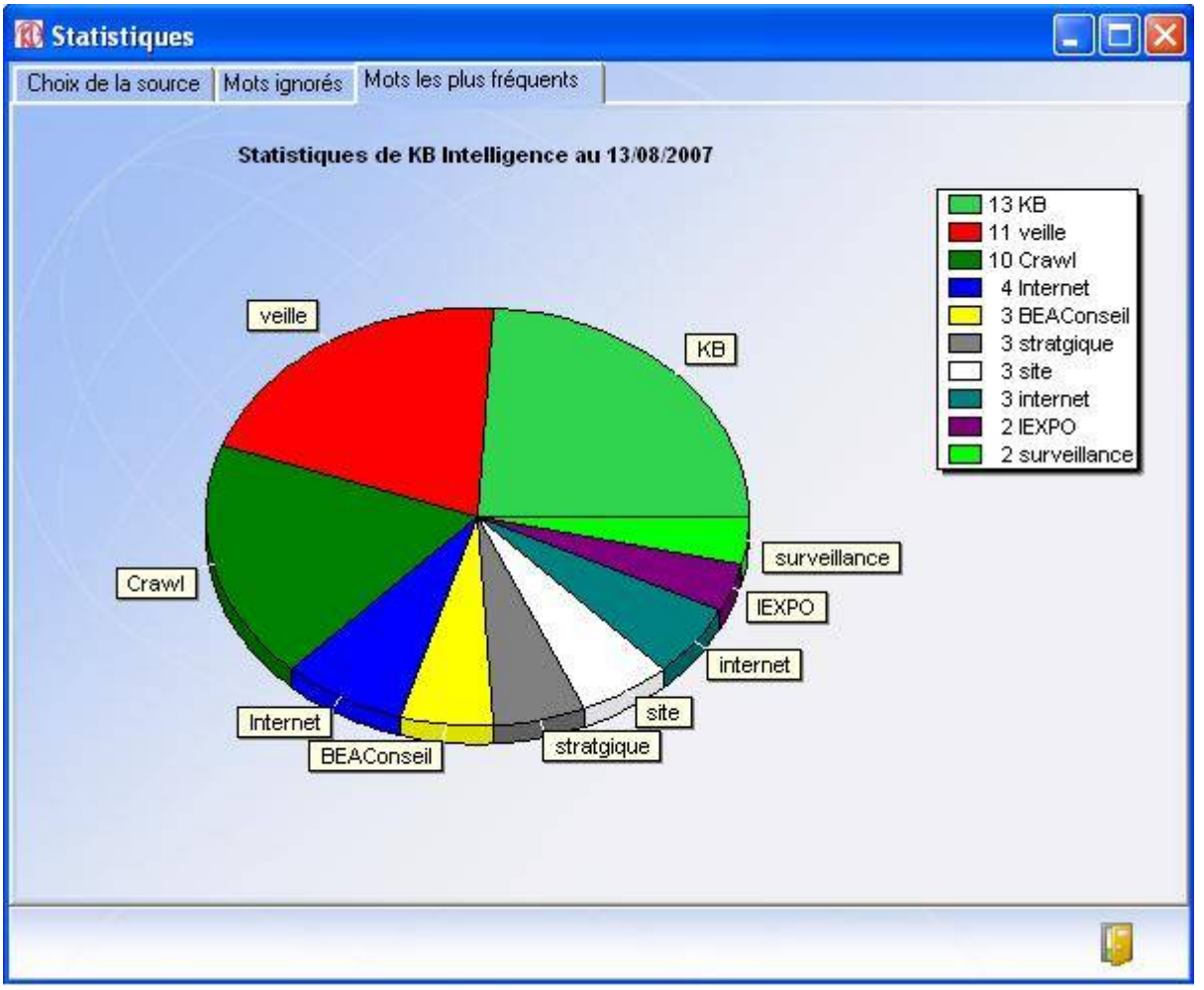
Note : Le nom de la source peut être omis, dans ce cas, le nom de la source est généré automatiquement. D'autre part, l'adresse URL peut ne pas contenir le nom du protocole « http:// » ; dans ce cas, il sera ajouté automatiquement.

13.5.5 Statistiques

Le module de statistiques permet, pour une source sélectionnée, de voir les 10 termes les plus souvent rencontrés dans les contenus de cette source.

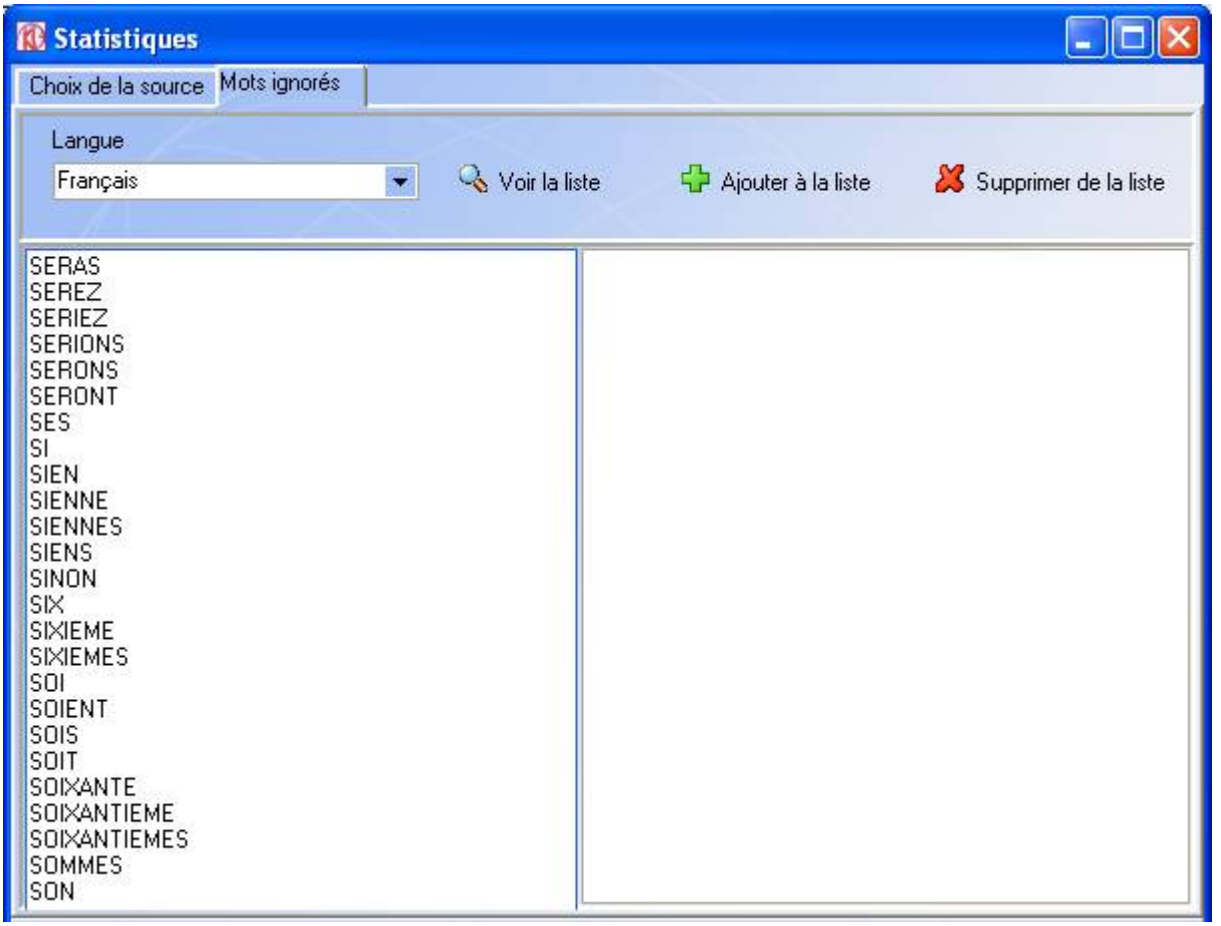


Sélectionner d'abord la source, puis la langue la plus utilisée dans les archives de cette source. Ensuite, cliquer sur le bouton « Mots les plus fréquents », patienter pendant le calcul puis un graphique de type « camembert » apparaît dans le troisième onglet :



Le camembert affiche des mots « parasites » comme « site ». On peut faire en sorte qu'il soit ignoré dans le calcul des statistiques pour se concentrer sur les mots intéressants.

Pour chaque langue répertoriée dans la base de données de KB Crawl (français, anglais, allemand, espagnol, italien, néerlandais, suédois), une liste de mots ignorés ou mots « noirs » est livrée avec le logiciel. Ces listes de mots ignorés sont stockées dans la base de données et entièrement paramétrables.



Il faut d'abord sélectionner dans la liste déroulante la langue voulue.

Pour ajouter des mots ignorés :

- saisir ces mots les uns à la suite des autres dans la zone de saisie située à droite de l'écran,
- cliquer sur le bouton « Ajouter à la liste».

Pour supprimer des mots ignorés :

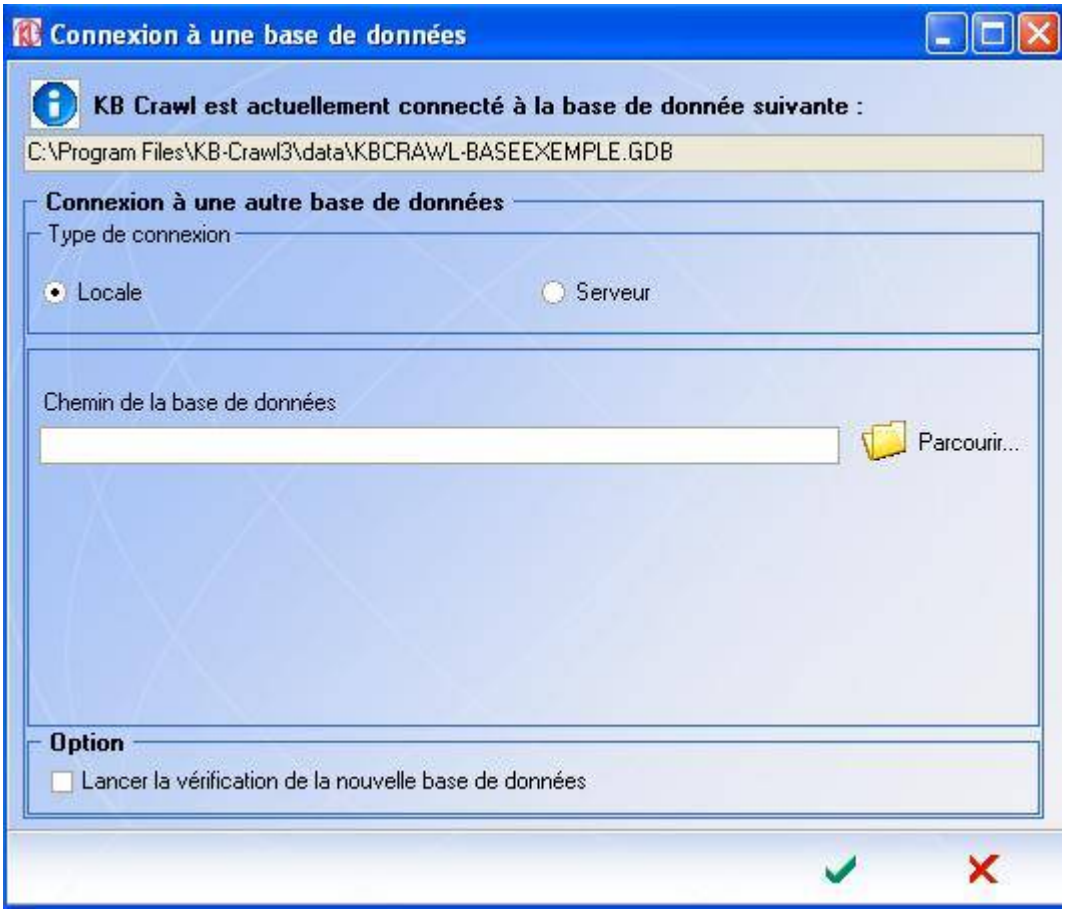
- saisir ces mots les uns à la suite des autres dans la zone de saisie située à droite de l'écran,
- cliquer sur le bouton « Supprimer de la liste».

On peut également afficher la liste des mots ignorés pour une langue donnée en cliquant sur le bouton « Voir la liste ».
La liste des mots ignorés s'affiche alors dans la zone située à gauche de l'écran.

13.6 Paramètres

13.6.1 Se connecter à une autre base de données

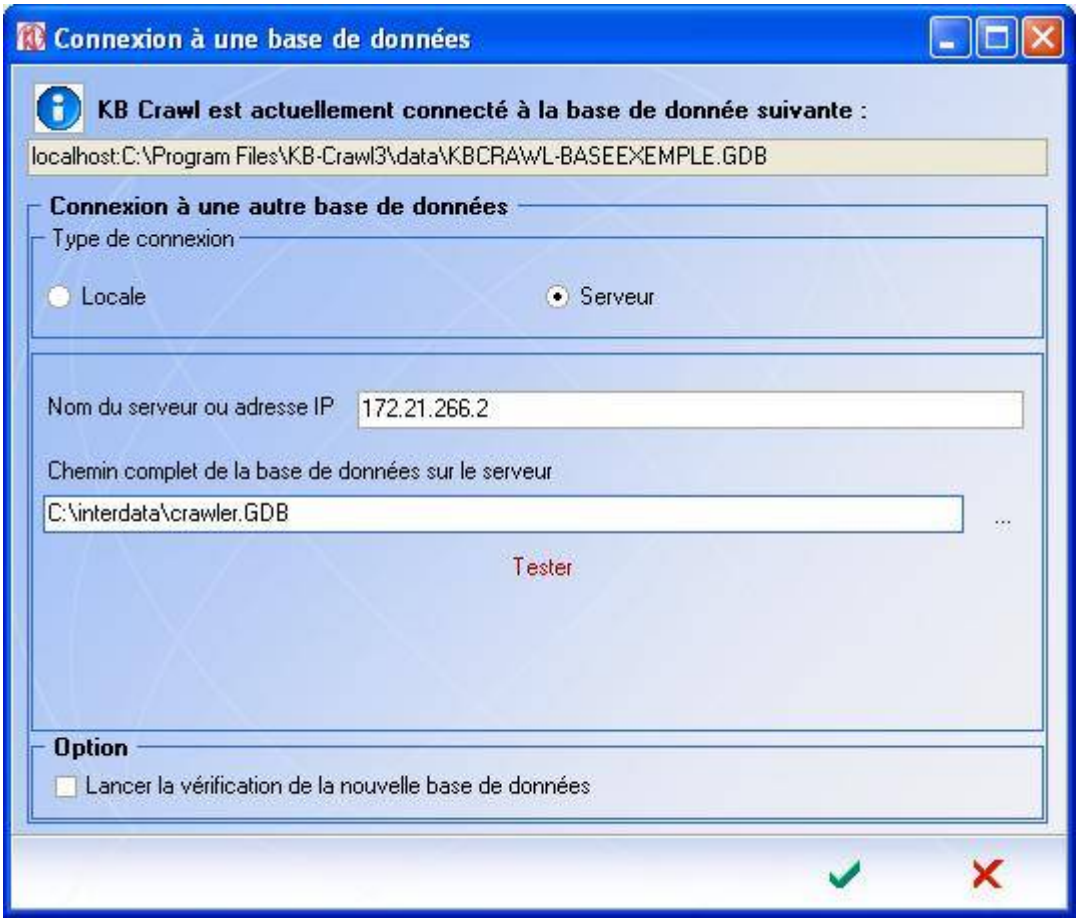
Il est possible de créer plusieurs bases de données et de passer de l'une à l'autre via KB Crawl.



La base de données peut être soit locale, soit distante (visible via le réseau LAN).
Cocher l'option correspondante (Locale ou Serveur).

Dans le cas où la base de données est locale à l'ordinateur sur lequel est installé KB Crawl, saisir le chemin entier du fichier GDB.

Dans le cas où la base de données est hébergée sur un autre ordinateur, saisir l'adresse IP ou le nom de cet ordinateur, puis le chemin complet de la base de données sur cet ordinateur.



Valider avec la coche verte pour effectuer la connexion.

En connexion de type serveur, un test de connexion préalable peut être effectué en cliquant sur le bouton « Tester »

En option, la base de données à laquelle KB Crawl va se connecter peut être vérifiée.

13.6.2 Grammaire du parser

La grammaire du parser est entièrement paramétrable. Elle peut donc être enrichie ou modifiée à volonté pour ajuster les performances de parsing.
KB Crawl fournit une grammaire par défaut la plus complète possible, mais celle-ci ne peut couvrir tous les cas de codes HTML et JAVASCRIPT rencontrés dans l'ensemble des pages du Web !
Il est donc parfois nécessaire d'enrichir cette grammaire afin de pouvoir récupérer certains liens et naviguer en profondeur de page en page dans un site Web.
A cet effet, la fonctionnalité d'export et d'import de la grammaire du parser permet de travailler à partir d'un fichier ASCII qui représente cette grammaire.

Chaque ligne du fichier correspond à une balise HTML et une variable ou une fonction que le parser « attend » à l'intérieur de celle-ci.
Une variable est suivie d'un signe « = » qui lui-même est suivi de guillemets ou de cotes qui entourent une valeur chaîne alors qu'une fonction est directement suivie (ou non) d'une

parenthèse ouvrante, elle-même suivie de valeurs chaines séparées par des virgules et délimitées par des guillemets ou cotes.

Chaque ligne est composée de 6 champs séparés par un « ; »

Champ 1 : libellé de la balise HTML ouvrante.
Champ 2 : libellé de la balise HTML fermante.
Champ 3 : libellé de la variable ou du nom de fonction à trouver entre les deux balises.
Champ 4 : la valeur de la variable ou les arguments de la fonction sont-ils à récupérer en tant que bloc de texte (O/N) ?
Champ 5 : la valeur de la variable ou les arguments de la fonction contiennent-ils des liens à parcourir (O/N) ?
Champ 6 : s'agit-il d'une variable ou d'une fonction ?

Exemple avec une variable :

href= « http://www.kbcrawl.net »

Ligne du fichier correspondante :

<A;>;HREF;N;O;N;

Balise ouvrante : <A
Balise fermante : >
Le contenu de la variable n'est pas à récupérer comme texte.
La variable contient une valeur chaine à interpréter comme un lien.
HREF n'est pas une fonction.
Remarque : Le parser n'est pas sensible à la casse des champs saisis.

<a href= « http://www.kbcrawl.net »>

Lorsque le parser rencontre la balise <A, il examine son contenu jusqu'à trouver « > ».
Lors de cet examen, il reconnait la variable « href » et se place après le signe « = » et extrait tout ce qui est entre guillemets ou entre cotes.

Exemple avec une fonction :

WINDOW.OPEN(« http://www.kbcrawl.net » , « Kbcrawl »)

Ligne du fichier correspondante :

<SCRIPT;</SCRIPT>;WINDOW.OPEN;N;O;O;

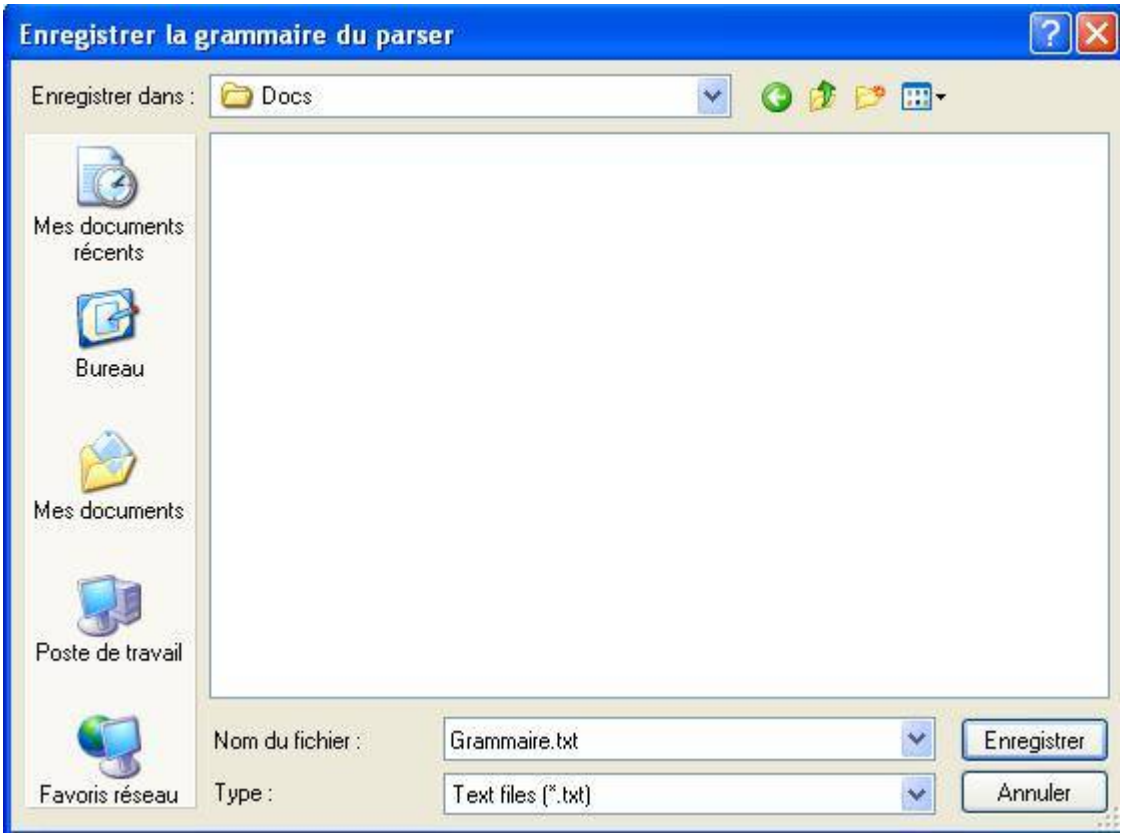
Balise ouvrante : <SCRIPT
Balise fermante : </SCRIPT
Nom de la fonction : WINDOW.OPEN
Les arguments de la fonction ne sont pas à récupérer comme texte.
Les arguments de la fonction sont à interpréter comme des liens.
WINDOW.OPEN est une fonction.

Lorsque le parser se trouve à l'intérieur d'une balise de script, s'il trouve WINDOW.OPEN, il sait que les arguments entre parenthèses qui suivent sont interprétables comme des liens.

Remarque : On ne peut pas savoir quels arguments parmi ceux appelés par la fonction sont des liens, donc par défaut, KB Crawl essaie de télécharger à partir de chaque lien potentiel. Ceux qui n'en sont pas donneront simplement lieu à une requête qui n'aboutit pas.

- Exporter la grammaire du parser

Exporte la grammaire du parser au format ASCII dans un répertoire désigné.



Entrer le nom du fichier à enregistrer puis cliquer sur « Enregistrer ».

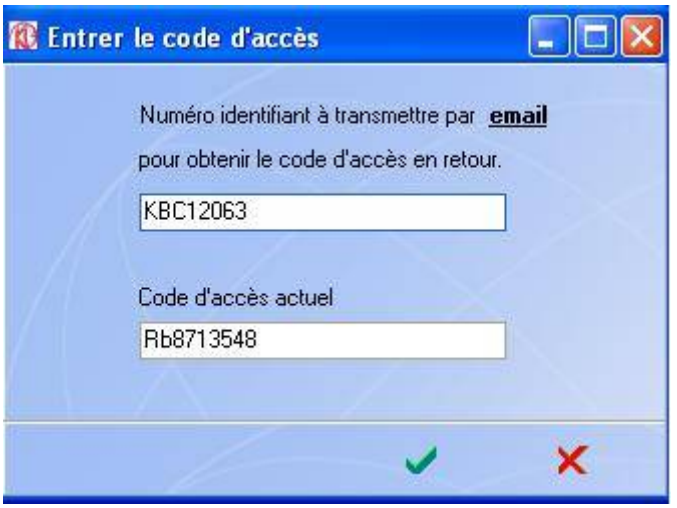
- Importer la grammaire du parser

Importe la grammaire du parser depuis un fichier désigné.

Sélectionner le fichier à importer puis cliquer sur « Ouvrir ».

13.6.3 Modifier la clé d'enregistrement KB Crawl

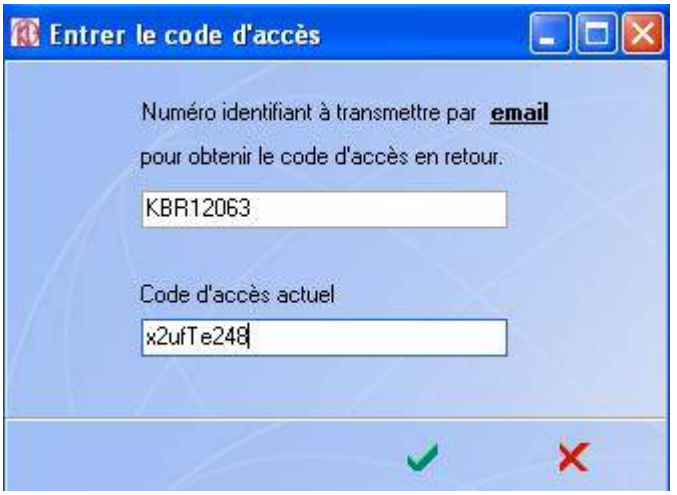
Il est possible de changer la clé d'utilisation de KB Crawl qui est inscrite dans la base de registre via l'interface de KB Crawl.



Cette fonctionnalité sera utile si l'on acquiert une clé d'utilisation définitive de KB Crawl. Il est ainsi aisé de passer d'une version de test de KB Crawl à une version définitive.

13.6.4 Modifier la clé d'enregistrement de KB Scraper

Il est aussi possible de modifier la clé d'enregistrement de KB Scraper via l'interface de KB Crawl.



Cette fonctionnalité sera utile si l'on acquiert une clé d'utilisation définitive de KB Scraper. Il est ainsi aisé de passer d'une version de test de KB Scraper à une version définitive.

13.7 Maintenance

13.7.1 Archives

13.7.1.1 Supprimer des éléments historiques

Cette opération de maintenance est utile, voire nécessaire, lorsqu'on gère une base d'archives de taille importante, elle permet de supprimer d'anciennes archives réduisant ainsi la taille de la base de données.

Tous les documents archivés antérieurement à la date choisie seront définitivement supprimés.

13.7.1.2 Optimiser la base d'archives

Lorsqu'un document est marqué comme supprimé, le comportement par défaut du module de gestion des archives est de le supprimer automatiquement de la base de données. (§ 3.7.4).
Il arrive que certains documents qui ne sont plus téléchargés figurent encore dans la base de données. Ces documents peuvent être supprimés afin d'optimiser la taille de la base de données.

13.7.2 Base de données

KB Crawl stocke toutes les informations acquises au cours des différents crawls ainsi que toutes les données liées aux paramétrages dans une base de données relationnelle Firebird.

Utiliser cette base de données présente un avantage majeur : bien que très performante et n'imposant pas de limite de volumétrie, elle ne requiert pas de maintenance contraignante qui nécessiterait l'intervention régulière d'un administrateur de base de données.

KB Crawl est équipé d'un nouveau système de maintenance automatique afin de dispenser de toute intervention manuelle.
Néanmoins, afin que les performances soient optimales en termes de rapidité ou d'espace occupé, et pour protéger la base de données d'éventuelles défaillances matérielles qui pourraient l'endommager, une série d'outils très simples à manipuler sont proposés dans le menu utilitaire.

13.7.2.1 Créer une sauvegarde compressée de la base de données

Toutes les informations liées à KB Crawl, qui constituent une véritable base de connaissances, sont stockées dans un seul et même fichier qui porte l'extension GDB.
Il est donc impératif de faire régulièrement une sauvegarde de ce fichier au cas ou des dommages subis par le disque dur en causeraient la perte ou la corruption irrémédiable.
Celui-ci contient principalement des documents extraits du Web, et peut donc être compressé afin de générer un fichier de sauvegarde dont la taille est optimisée.

Figure 84 : Sauvegarde de la base de données.

Dans un premier temps, définir le nom et l'emplacement du fichier de sauvegarde (on a l'habitude d'utiliser l'extension GBK pour le fichier de sauvegarde mais il n'y a aucune obligation) puis cliquer sur le bouton « Lancer la sauvegarde ».

Toutes les actions liées à la sauvegarde réalisées par l'utilitaire GBAK fourni avec la base de données Firebird sont monitorées dans la fenêtre de sauvegarde.

Figure 85 : Journal de la sauvegarde.

Les dernières lignes du journal de sauvegarde témoignent du bon achèvement du processus et la taille (en octets) du fichier de sauvegarde créé est indiquée.

13.7.2.2 Restaurer à partir d'une sauvegarde compressée

Toutes les informations ainsi que la structure de la base de données sont stockées dans le fichier de sauvegarde et peuvent être « remontées » (ou restaurées) à tout moment pour reconstituer une base de données utilisable par KB Crawl.

Il suffit pour cela de désigner un fichier de sauvegarde et de cliquer sur le bouton « Restaurer » :

Figure 86 : Restauration d'une sauvegarde.

Un nouveau fichier de base de données Firebird à l'extension GDB est créé à côté du fichier de sauvegarde portant le même nom que le fichier de sauvegarde comme préfixe et une série de chiffres issus de la date et l'heure exacte de la création de la nouvelle base de données.

Un message de confirmation précisant ces informations apparaît ; il est possible de modifier le nom de la base de données

La confirmation est suivie de la restauration de la sauvegarde, elle aussi journalisée :

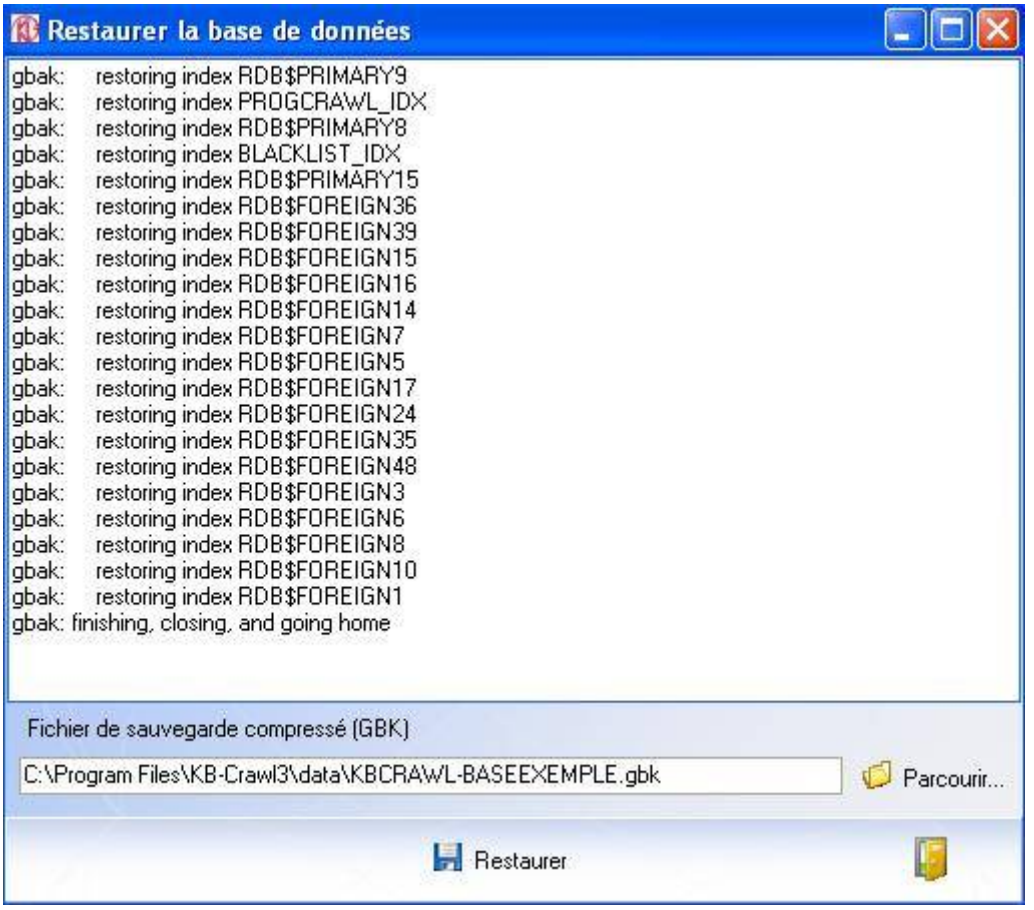


Figure 87 : Journal de la restauration d'une sauvegarde.

13.7.2.3 Compacter la base (=réduire la taille de la base)

La base de données Firebird grossit en taille régulièrement : à chaque fois qu'un crawl est lancé, une ou plusieurs pages peuvent être archivées, soit parce qu'il s'agit d'un crawl d'initialisation, ou bien parce qu'une alerte sur un document téléchargé justifie son stockage dans la table d'archives.

A chaque fois qu'un document est stocké dans la base de données, celle-ci s'alloue un espace nécessaire à ce nouveau stockage sur le disque dur. Autrement dit, à chaque fois qu'un document est stocké dans la base de données, la taille de cette dernière s'accroit d'un nombre d'octets supérieur ou égal à la taille du document.

En revanche, lorsqu'un document est supprimé de la base de données, parce qu'une nouvelle version du document vient le « chasser » de la table d'archive (§ 1.7.2), ou tout simplement parce qu'une source est supprimée, les informations sont effacées mais l'espace qui lui était réservé dans la base demeure, ceci pour des raisons techniques liées à l'optimisation des performances.

Ainsi, à force d'ajouter et de supprimer des documents, la base de données comporte de nombreux espaces physiquement accaparés mais inutilisés. L'espace qu'elle occupe sur le disque dur peut donc être optimisé.
C'est ce qu'on appelle un compactage.

A chaque ouverture de KB Crawl, celui-ci confronte le nombre de documents stockés dans la base de données à sa taille réelle sur le disque dur et propose, si cela est pertinent, d'effectuer un compactage.
Cette fonctionnalité est également accessible manuellement.
Lors de l'accès à cette fonctionnalité, un message de confirmation apparaît et avertit que l'opération prend un temps relatif à la taille de la base. Ce temps dépend également de la vitesse du processeur et d'autres paramètres, ce qui nous empêche de donner un ordre de grandeur.

Pour des bases de données très volumineuses (plusieurs giga octets), plusieurs heures de traitement peuvent être nécessaires, tout comme la défragmentation d'un disque dur qui est une opération comparable.

A la fin de l'opération, un message apparaît et informe de la nouvelle taille de la base de données :

Remarque : Le compactage revient à faire une sauvegarde de la base de données puis une restauration de la sauvegarde de la base de données en écrasant la base de départ.

13.7.2.4 Vérifier la base de données

Lors de défaillances mécaniques du disque dur, ou parce qu'une utilisation inadaptée est faite de la base de données (exemple : ordinateur éteint en cours de traitement), la base de données peut subir des dommages plus ou moins importants et ainsi se dégrader partiellement. On dit alors que la base de données est « corrompue ». La plupart des corruptions de la base de données passent inaperçues parce qu'elles sont tolérées par le serveur de base de données Firebird.

Lorsqu'un blocage lié à une corruption de la base de données survient, celle-ci est passée par des stades intermédiaires.
Pour éviter cela, un système de vérification automatique de la base est mis en place à chaque ouverture de KB Crawl. Il est nécessaire d'attendre que la vérification soit terminée avant l'ouverture de KB Crawl.

Il est possible de désactiver la vérification automatique de la base de données en appelant l'application KB Crawl avec le paramètre « NOCHECK ».
Ceci est fortement déconseillé pour les raisons évoquées plus haut et revient à assumer les risques encourus, qui consistent dans le pire des scénarios à perdre définitivement les données stockées.

13.7.2.5 Vérifier la taille de la base

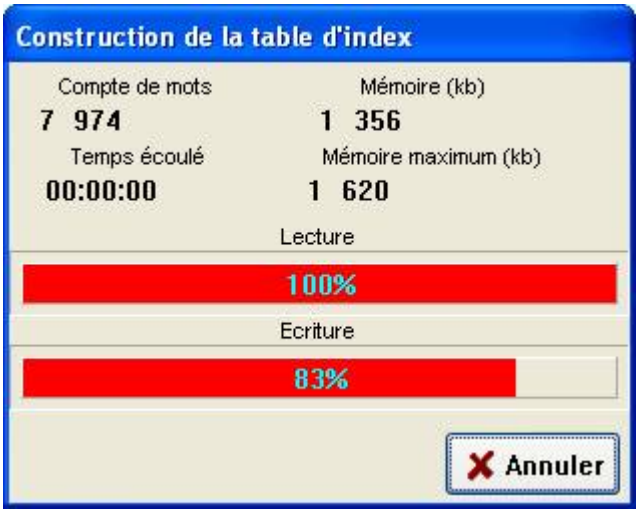
Cette fonctionnalité est utilisée à chaque ouverture de KB Crawl pour contrôler la taille de la base de données et voir si celle-ci ne peut être optimisée.
Cependant, le contrôle peut être fait à tout moment, ce qui peut être utile dans le cas de serveurs sur lesquels on ne ferme pas souvent l'application.

13.7.3 Service d'indexation

13.7.3.1 Vérifier la synchronisation de l'indexation et des archives

Le moteur d'indexation des contenus de KB Crawl stocke tous les mots indexés dans la base de données de KB Crawl, plus précisément dans une table d'index. Il tient à jour également cette table d'index au fur et à mesure des crawls si l'indexation automatique est activée.

Pour effectuer une quelconque recherche avec le moteur de recherche de KB Crawl ou pour utiliser les alertes avancées, il est indispensable que la table d'index soit à jour, c'est-à-dire parfaitement synchrone avec les contenus archivés dans la base de données.
Si l'indexation automatique n'était pas utilisée et qu'on l'active parce que l'on souhaite utiliser les alertes avancées ou que l'on souhaite effectuer une recherche, la table d'index doit être reconstruite pour être synchronisée avec les archives.
Cette opération peut prendre plusieurs minutes si la base d'archives est volumineuse.



14 Glossaire

ADSL
(*Anglais : Asymmetrical Digital Subscriber Line*).

(Français : Ligne asymétrique numérique)
Technologie capable de transporter plusieurs mégabits par seconde sur les deux fils de cuivre du téléphone. Les données peuvent être transmises jusqu'à 8 Mbits/s en téléchargement. Cette technologie est dite asymétrique car le débit descendant (download) est différent du débit montant (upload).

BASE DE DONNEES
(Anglais : database)
Fichier ou ensemble de fichiers disque ou mémoire permettant le stockage permanent ou temporaire et l'accès à des informations structurées.

BROWSER
(Français : navigateur).
Programme utilisé pour explorer le Web. Les deux principaux browsers du marché : Internet Explorer (Microsoft) et Firefox. Les Québécois disent volontiers fureteur ou butineur.

SOURCE
Une source est un ensemble de pages Web dont le contenu textuel a été rassemblé puis stocké dans la base de données de KB Crawl.
Il se définit principalement par son point d'entrée (ou adresse de départ) et une profondeur de page et de site.

CRAWLER
(Français : Araignée)
C'est la partie d'un moteur de recherche qui surfe sur la toile, enregistre les URL, classe les mots-clés et le texte de chaque page qu'il trouve. En français, un terme souvent employé est robot. Les synonymes employés en anglais sont aussi « bots » et « spiders »

HTML
(Anglais : HyperText Mark-up Language).
Langage de description des pages Web dérivé du SGML. Il est composé d'une suite de signes ASCII, dans laquelle sont inclues les commandes spéciales concernant le formatage des pages, la police de caractères et les multimédia.

HTTP
(Anglais : HyperText Transfer Protocol).
Méthode utilisée pour transporter des pages HTML du WWW sur le réseau. L'accès aux services Web se fait en donnant une adresse de type http://nom de domaine/répertoire....

INTERNAUTE
Utilisateur de l'internet.
Note : On rencontre aussi le terme « cybernaute ».
(Journal officiel du 16 mars 1999 "Vocabulaire de l'informatique et de l'internet")

INTERNET
Ensemble de réseaux de toutes tailles interconnectés par le protocole IP. Le point de départ d'Internet fut ARPANet, c'est à dire un réseau de quatre ordinateurs que relièrent des scientifiques du ministère de la défense américaine en 1969. Dans les années qui suivirent, de plus en plus d'universités et d'instituts de recherche se sont joints à eux.

LAN
(acr. angl.)
(Anglais : Local Area Network)
Réseau local : Réseau situé dans une zone réduite ou dans un environnement commun, tels qu'un immeuble ou un bloc d'immeubles. Un réseau local devient une partie d'un réseau étendu lorsqu'une liaison est établie (via des modems, routeurs distants, lignes téléphoniques, satellites ou une connexion hertzienne) avec un gros système, un réseau de données public (Internet par exemple) ou un autre réseau local.

PARSING
Analyse syntaxique ou analyse grammaticale d'un document informatique (ex: HTML, XML, etc.).

PROVIDER
(ou Access Provider)
Fournisseur d'accès à l'Internet.

SOCKET
Deux processus indépendants sur deux machines distinctes, communiquent entre eux via les sockets.

URL
(Anglais : Uniform Resource Locator).
Adresse Internet exploitée par les navigateurs (Internet Explorer ou Firefox, par exemple). C'est l'adressage standard de n'importe quel document, sur n'importe quel ordinateur en local ou sur Internet.
Structure de base d'une URL :
protocole://serveur/répertoire/document.extension
http://www.yahoo.fr

WEB
Le Web (ou toile pour les Canadiens) est l'abréviation utilisée pour désigner le World Wide Web (le www des URL).
C'est un concept développé par les chercheurs du CERN, dont Tim Berner-Lee qui permet de rendre accessible, via le réseau Internet, des collections de pages hébergées sur des millions de serveurs répartis dans le monde.