Fiche n°1 :
Construire un corpus avec Gromoteur

Sommaire

Préambule :

On commencera par démarrer son ordinateur sous Windows

Pour la session de prise en main de Gromoteur présentée dans cette fiche, on fait le choix de travailler sous Windows.

Il sera possible par la suite de travailler sous d'autres systèmes d'exploitation (Linux, MacOsX).

Etape n°1 : Récupérer Gromoteur

Gromoteur a un site web : http://gromoteur.ilpga.fr/

Lancez votre navigateur et allez sur la page web de Gromoteur.

Sur la page précédente, localisez cette zone :

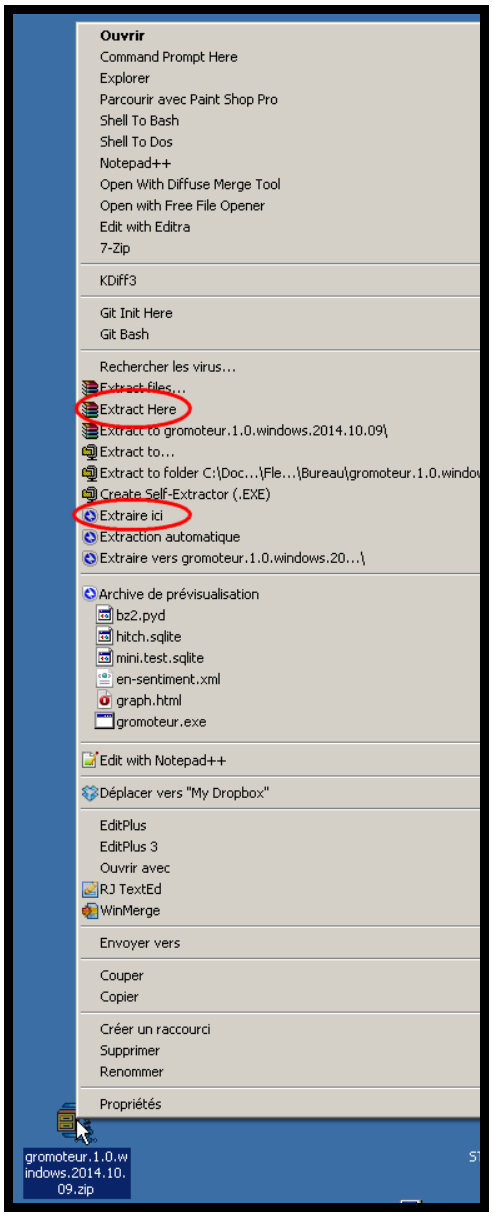
Faire un clic-droit sur le lien, et enregistrez la cible sur le bureau de votre ordinateur

(cf figure ci-dessous)

Etape n°2 : Dézipper l'archive téléchargée

1. Localisez l'archive téléchargée :

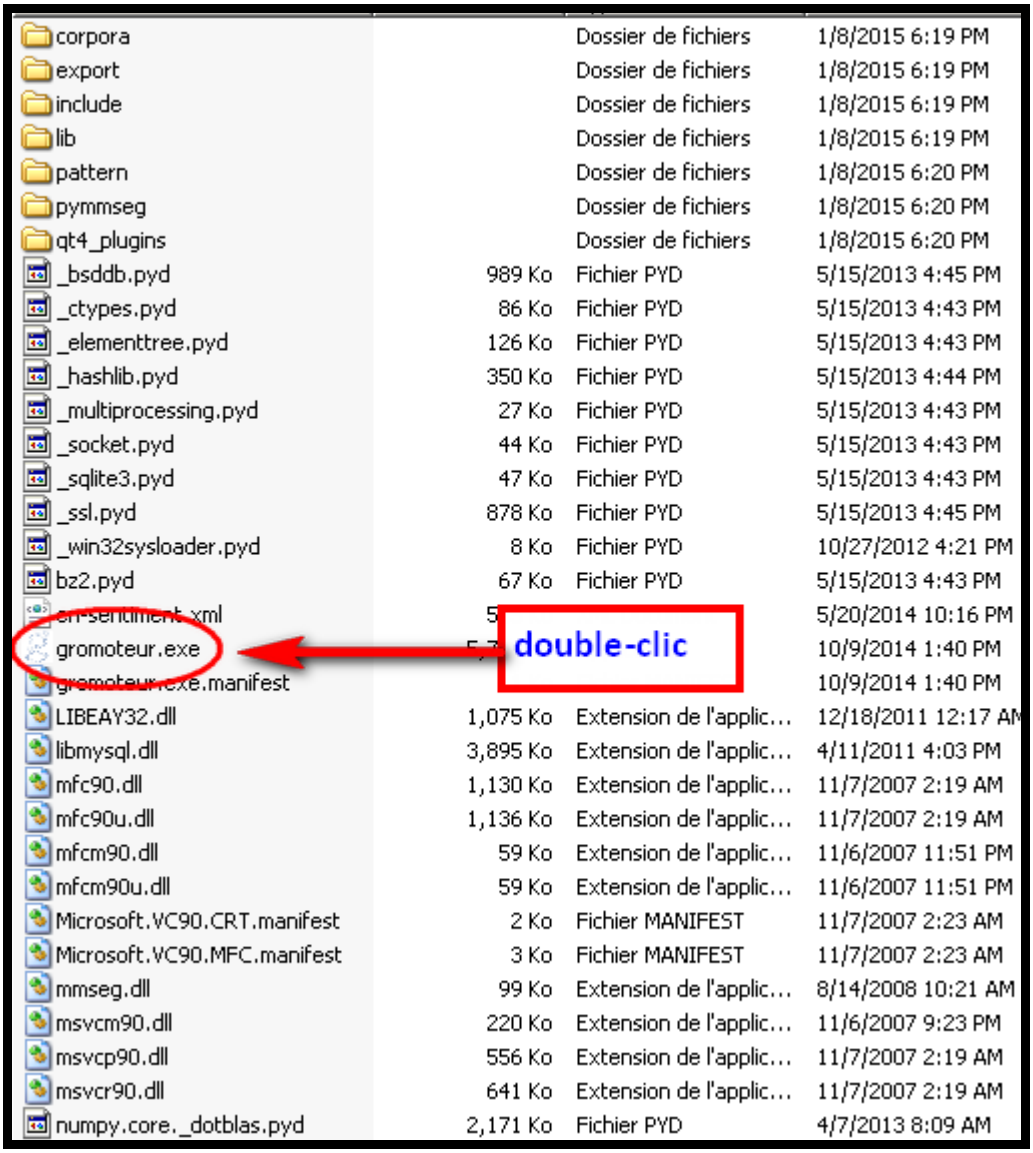
2. Dézippez-la : faire un clic-droit sur l'archive et repérer l'instruction « Extraire ici » ou « extract here » :



Etape n°3 : Démarrer Gromoteur

Une fois « dézippée », un dossier apparaît sur le bureau :

Pour lancer Gromoteur, ouvrir le dossier créé et localisez le programme **gromoteur.exe** :



Attention : suivant la configuration de windows, il est possible que l'extension .exe ne soit pas visible. Repérer l'icône du programme.

Etape n°4 : Récupérer la liste des URLs à traiter

Le fichier à traiter est disponible sur iCampus à cette adresse :

http://www.tal.univ-paris3.fr/trameur/FICHES/liste-url-newspaper.txt

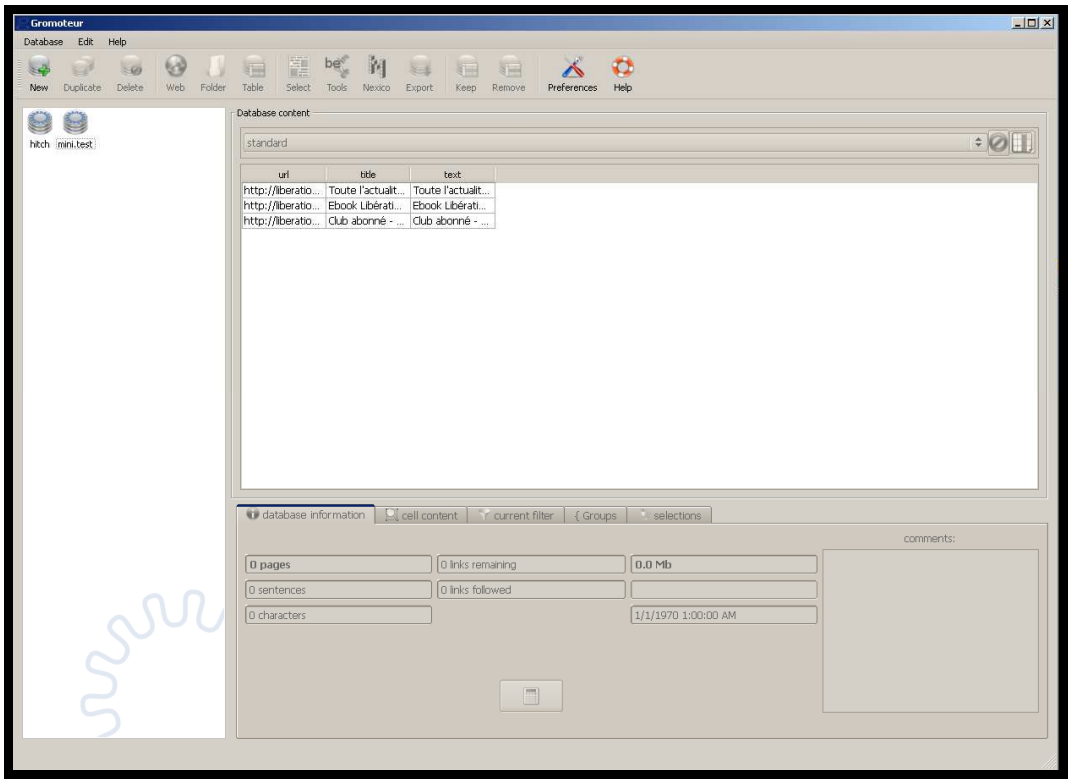
Ce fichier contient des adresses des pages d'accueil de journaux en ligne (nationaux ou régionaux) : il contient 67 adresses.

On en voit ci-dessous un extrait :

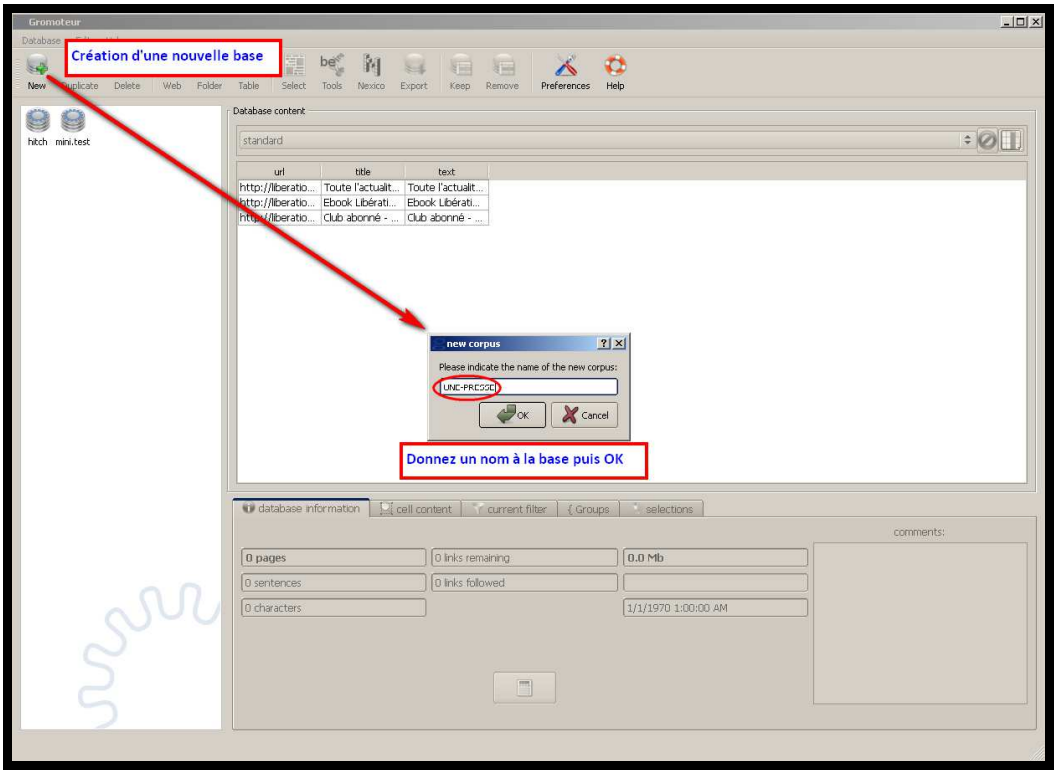
Enregistrer ce fichier sur le bureau de votre ordinateur, par exemple à côté des ressources déjà disponibles :

Etape n°5 : Paramétrer Gromoteur

Une fois lancé, Gromoteur apparaît !

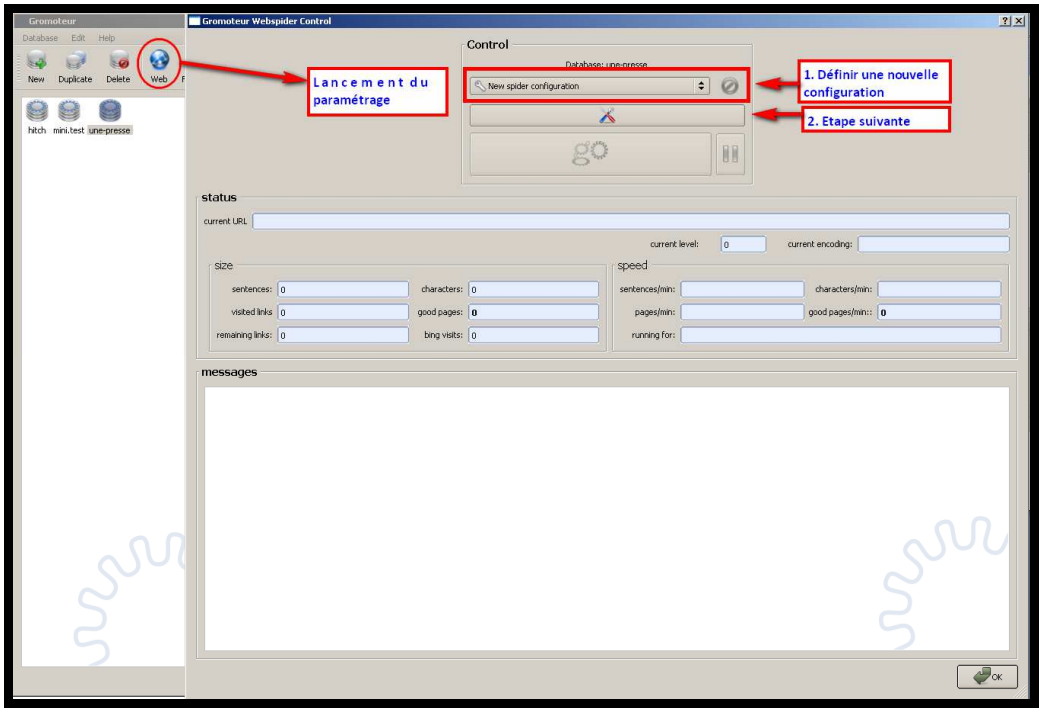


Créer une nouvelle base :

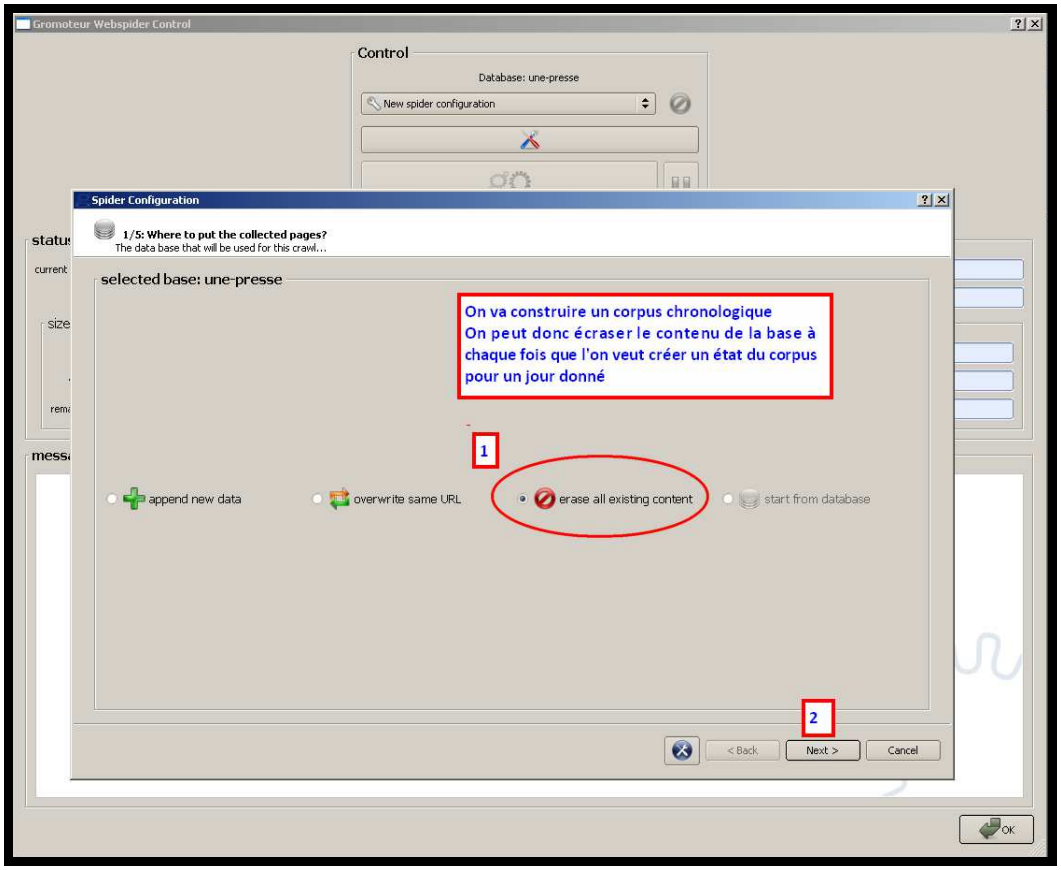


La nouvelle base apparaît : sélectionnez-la

Lancez le paramétrage :



Dans ce premier onglet, on choisit de définir une nouvelle configuration pour l'aspiration à mettre en œuvre.



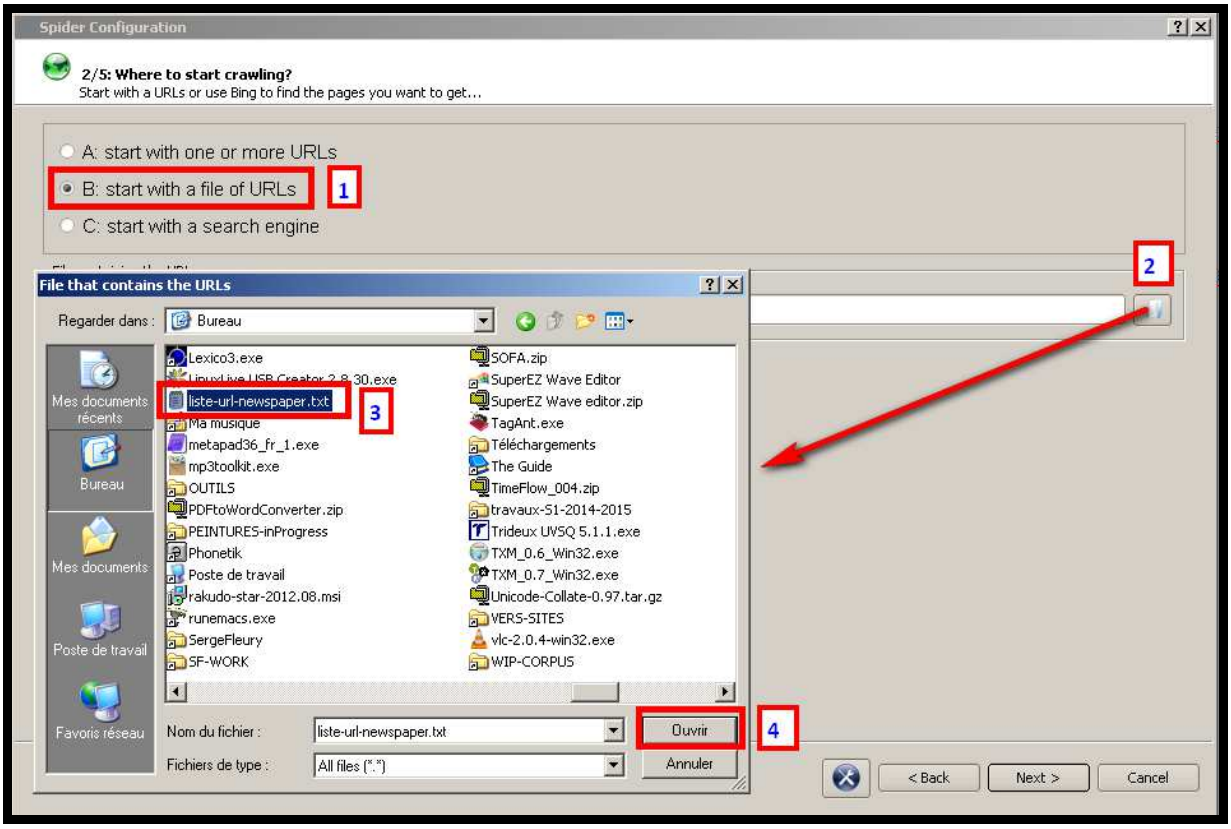
Dans cet onglet, on choisit la manière dont les données seront stockées dans la base.

On peut choisir de concaténer les données à celles déjà disponibles dans la base : cela signifie que si on lance plusieurs fois le programme d'aspiration , la base concatènera toutes les aspirations.

On peut aussi choisir de vider la base à chaque fois que l'on relance une nouvelle aspiration : dans ce cas il est peut-être prudent de sauvegarder les données stockées après chaque aspiration si on souhaite les réutiliser plus tard .

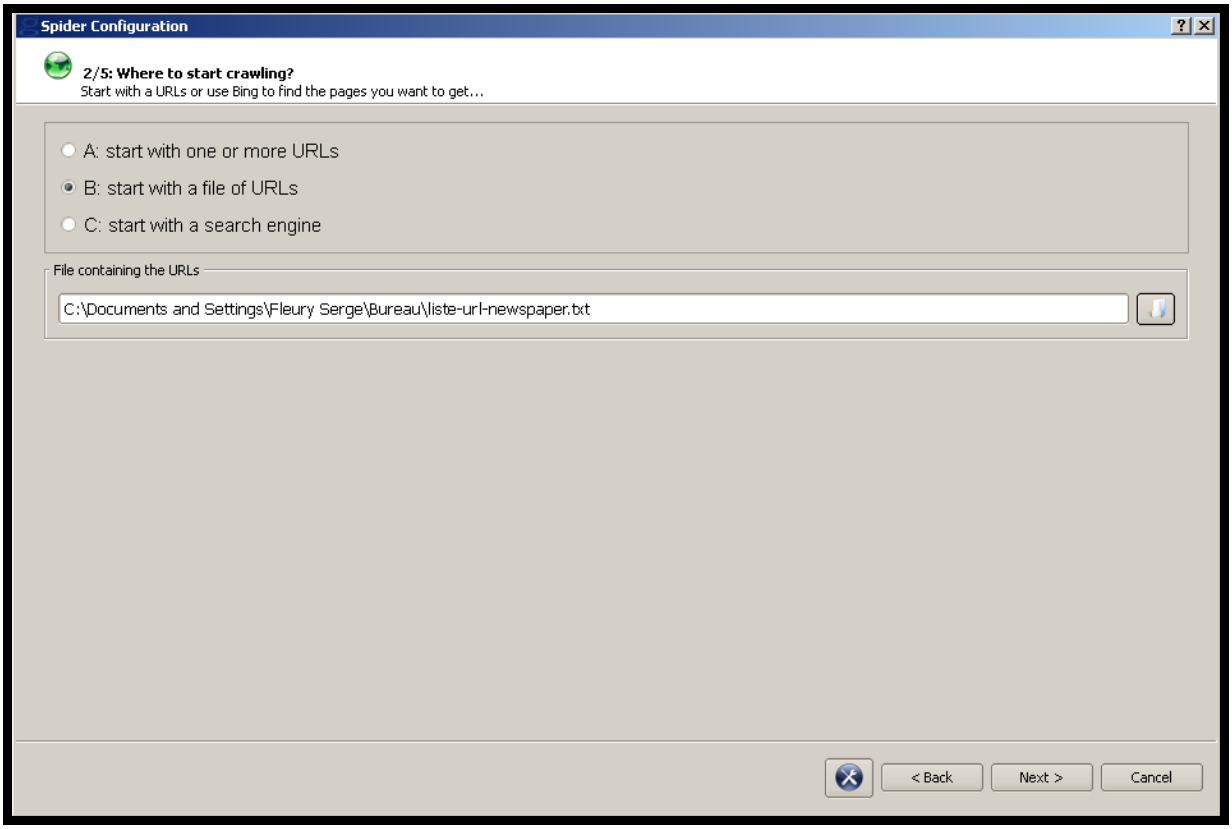
IMPORTANT :

Pour cette première session, on choisit ici d'écraser le contenu de la base (qui est vide puisqu'on vient de la créer), on prendra soin de sauvegarder le résultat produit (cf étape Export plus loin) avant de relancer une nouvelle aspiration.

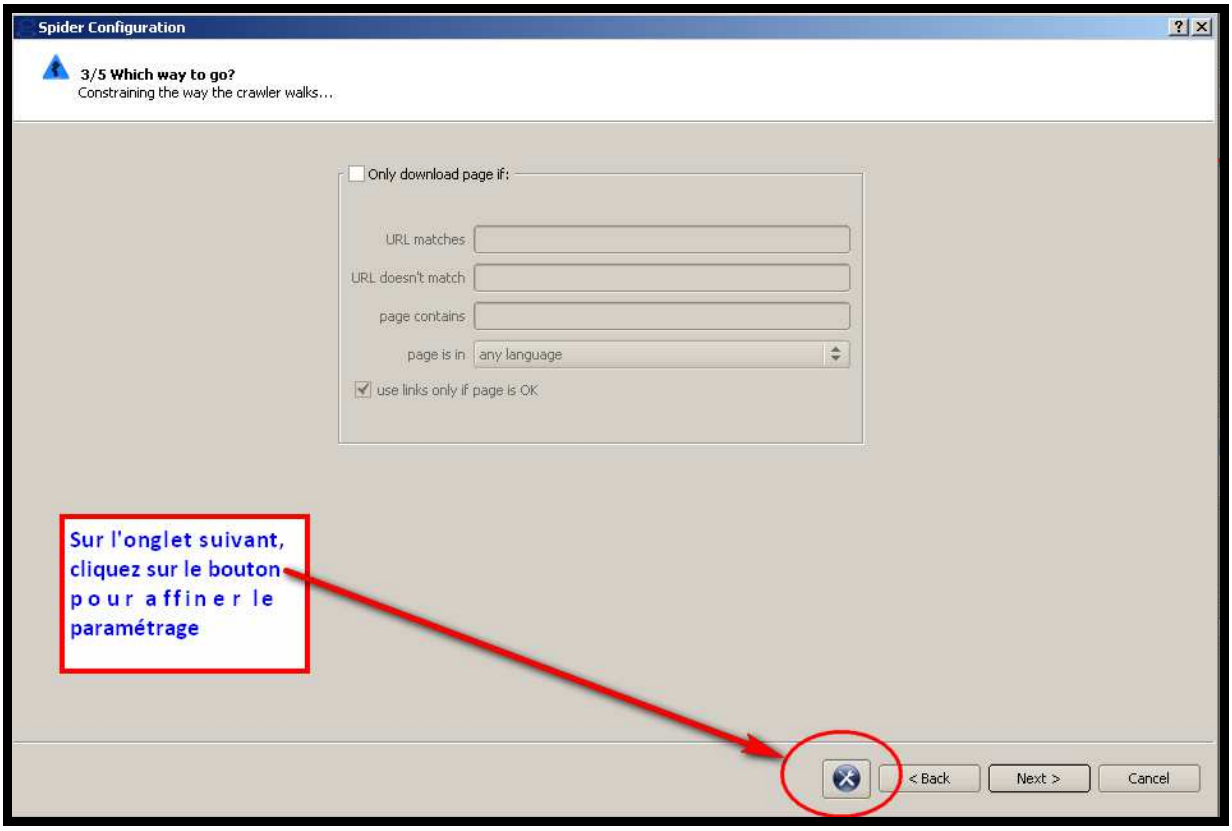


Dans cet onglet, on choisit de paramétrer l'aspiration en sélectionnant un fichier contenant les différentes adresses visées.

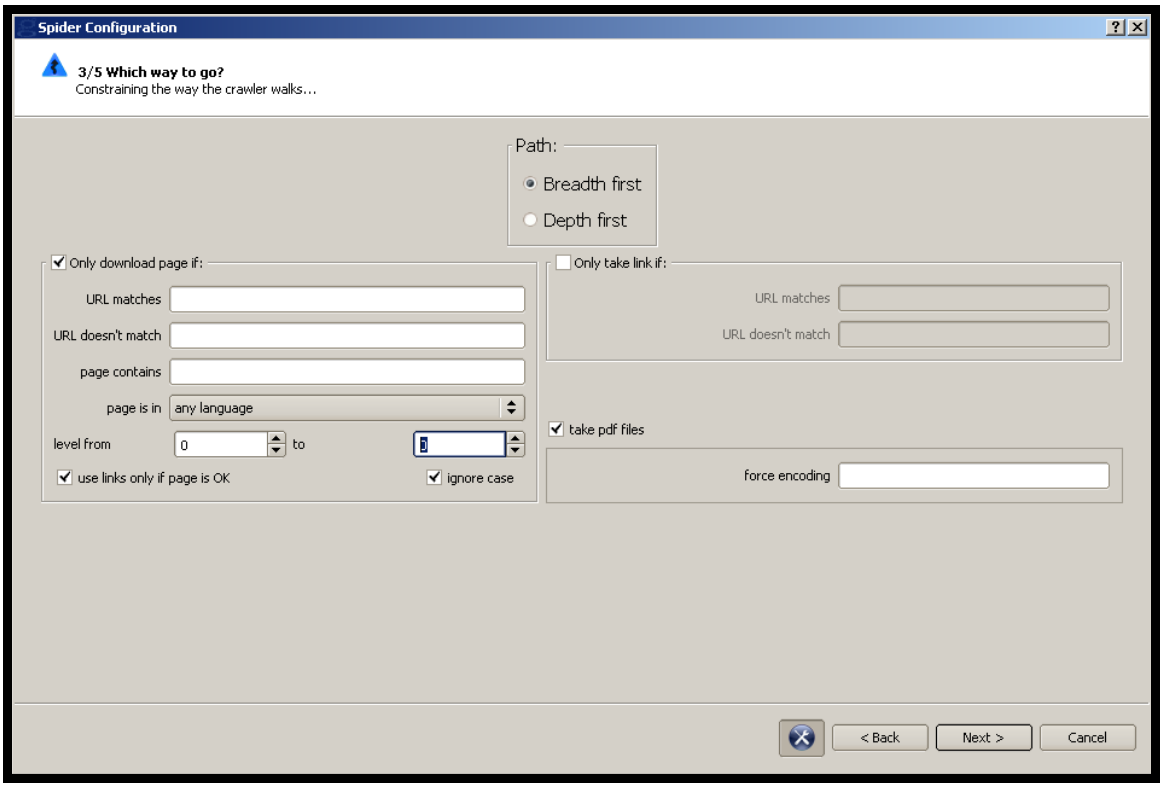
On indique donc ici le fichier récupéré tout à l'heure situé sur le bureau.



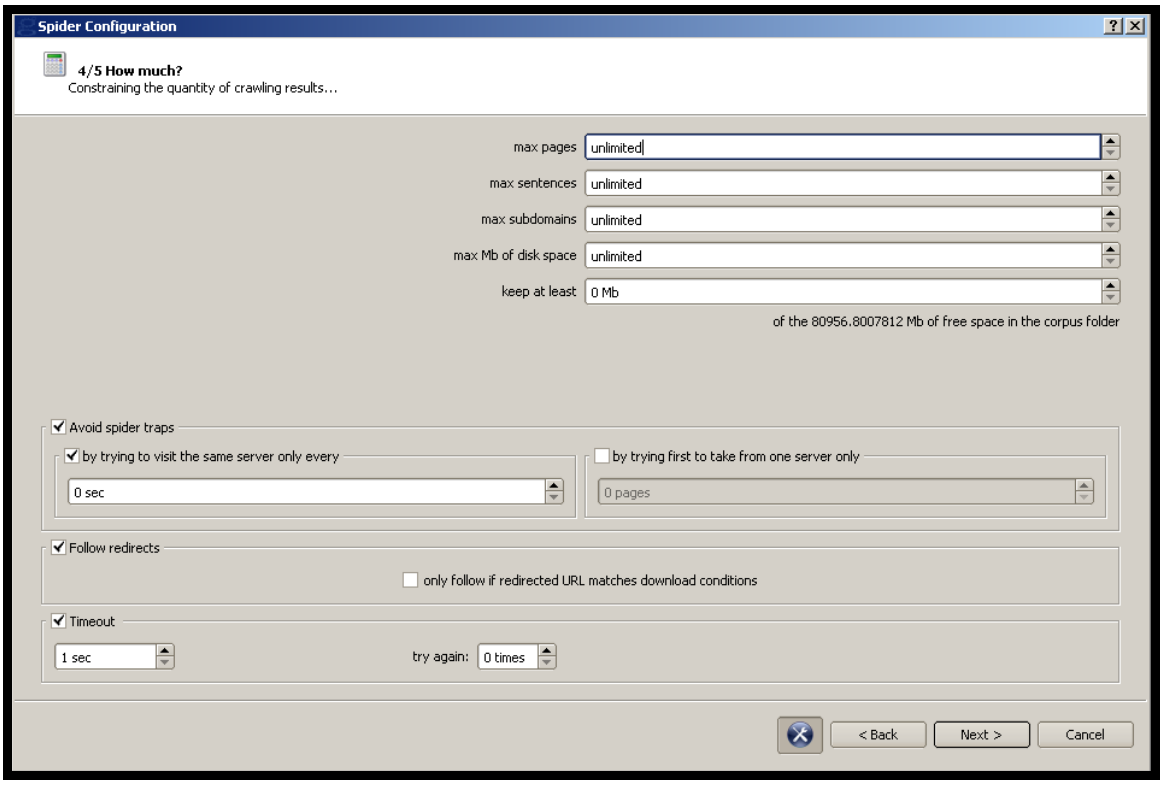
Après cette étape, tapez sur Next



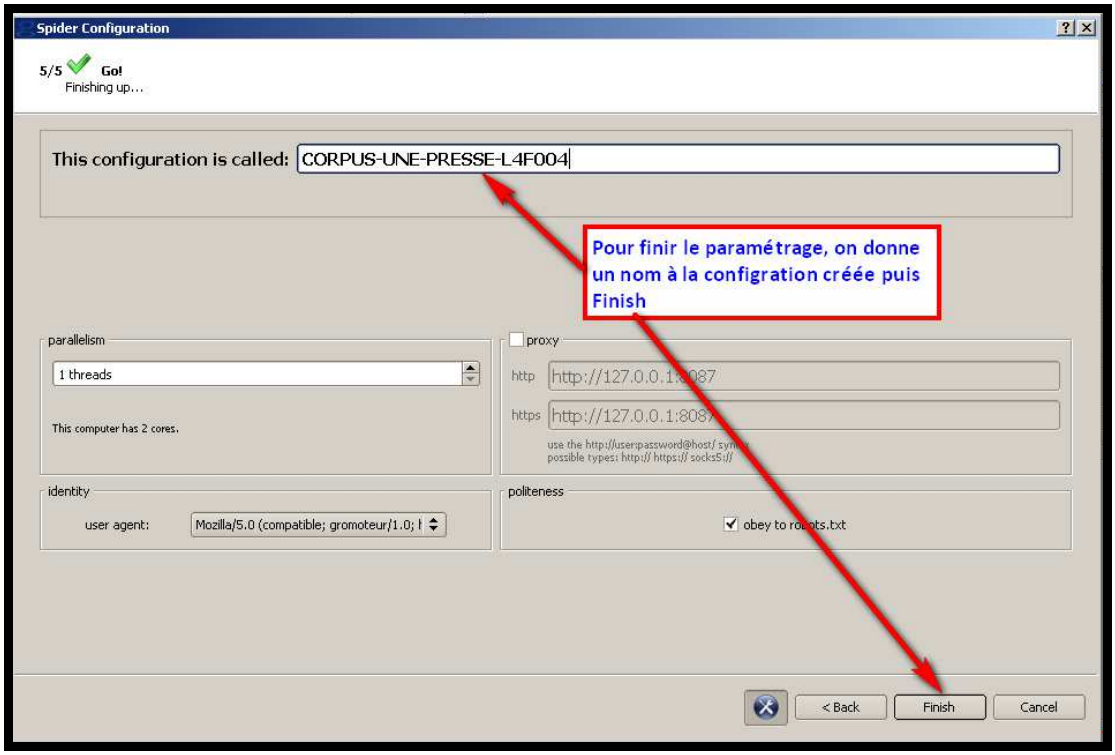
Paramétrez l'onglet suivant comme ceci :



Après cette étape, tapez sur Next. On laisse l'onglet suivant tel quel :

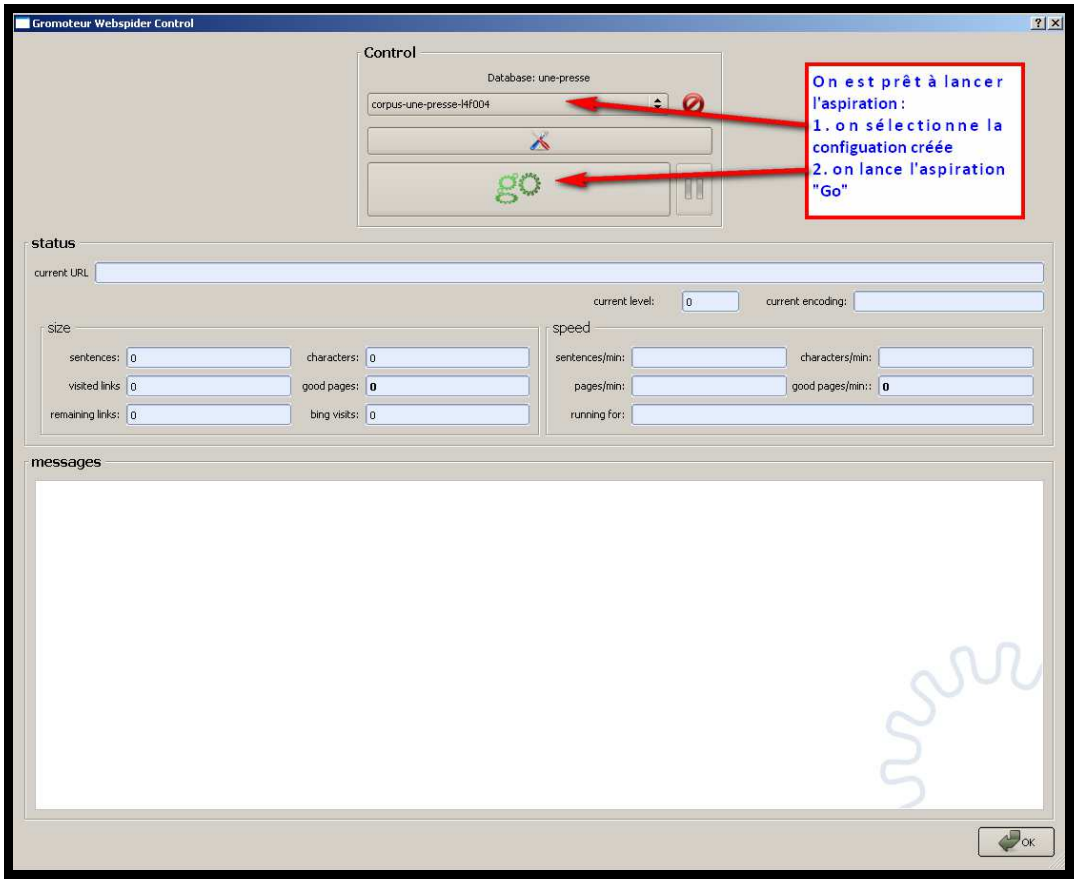


Après cette étape, tapez sur Next



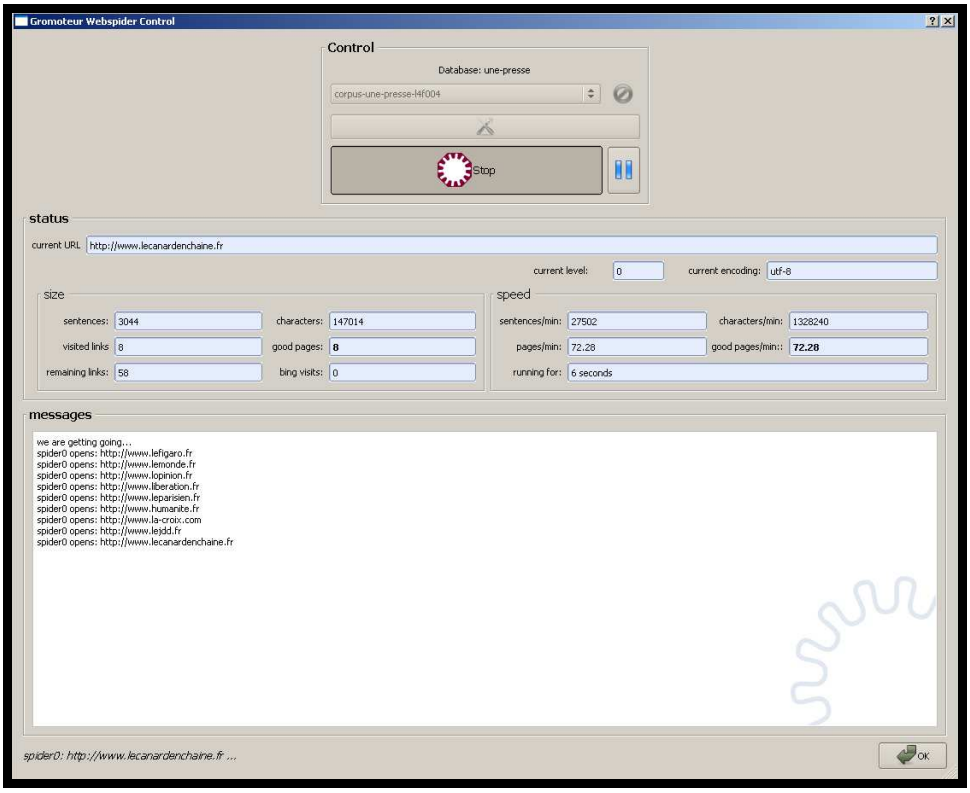
L'onglet précédent est le dernier onglet de paramétrage.

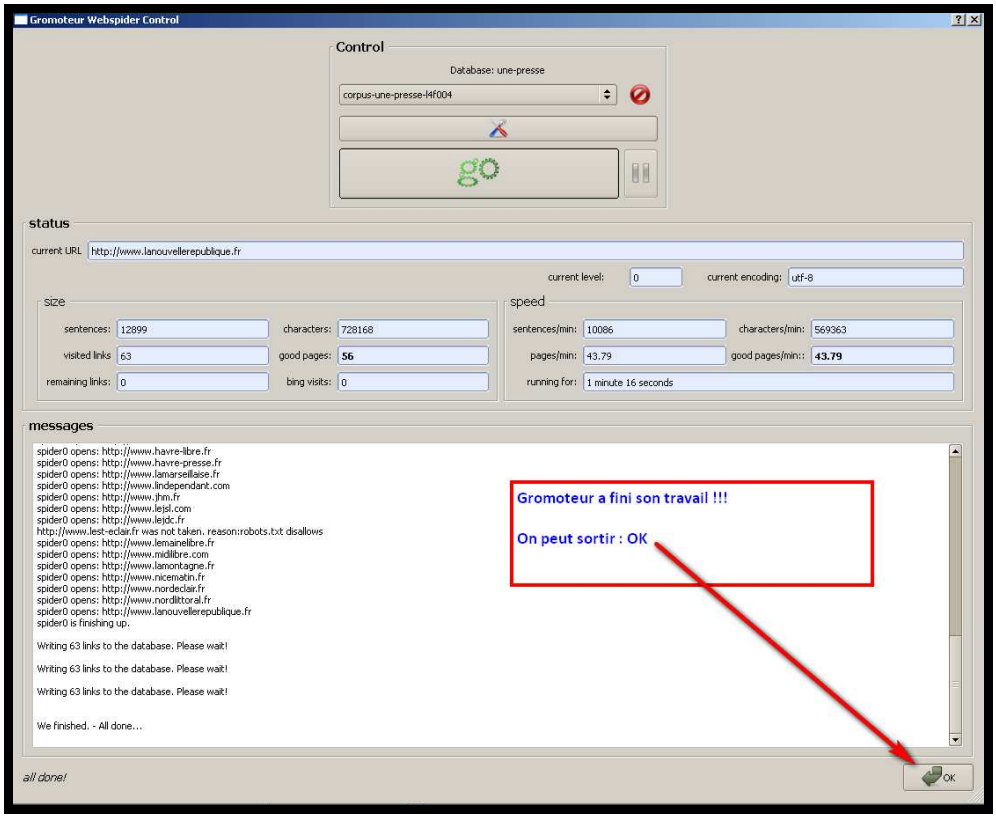
On est prêt à lancer l'aspiration des pages visées :



Etape n°6 : Lancer l' «aspiration »

Gromoteur démarre, on touche à rien !!!



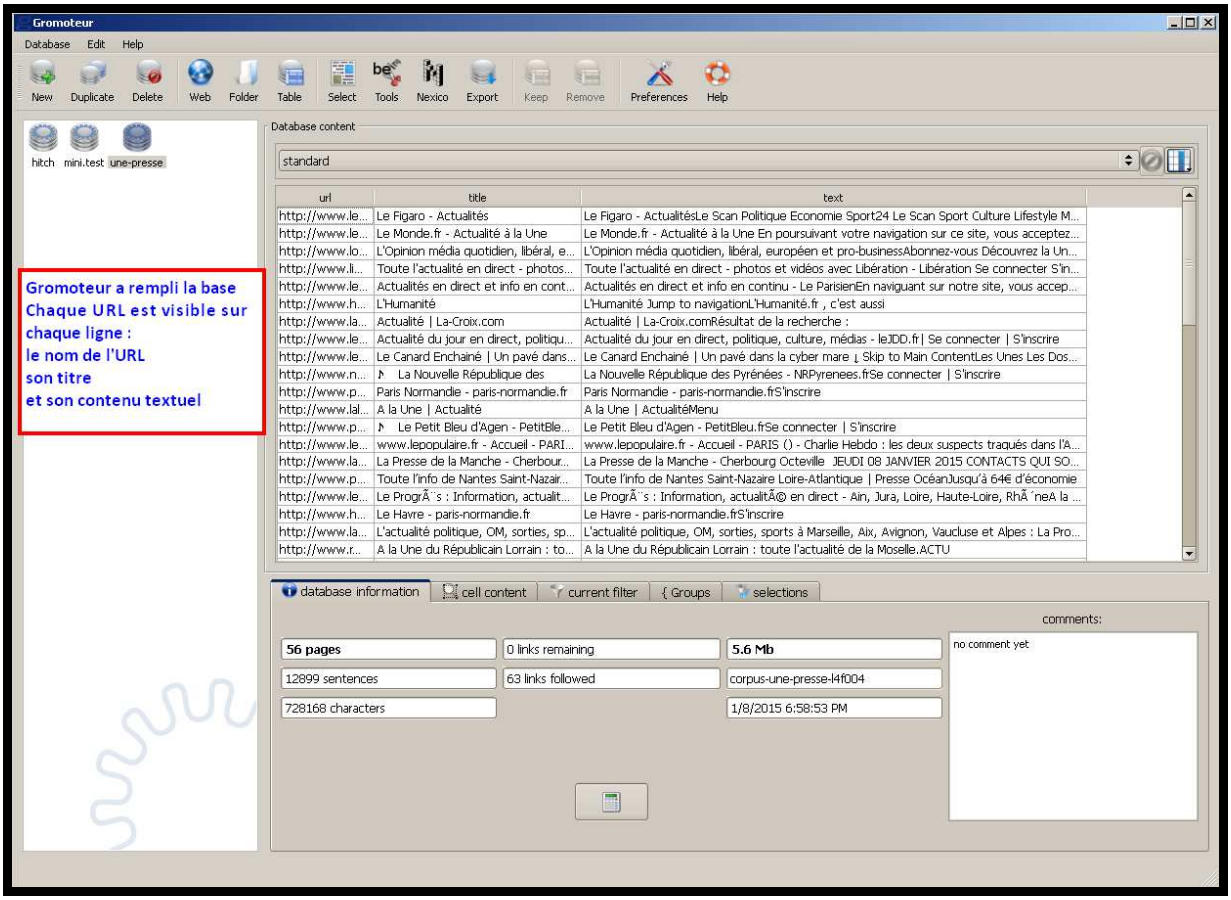


Gromoteur a finit son travail (All done !).

Etape n°7 : Vérifier le résultat produit par Gromoteur

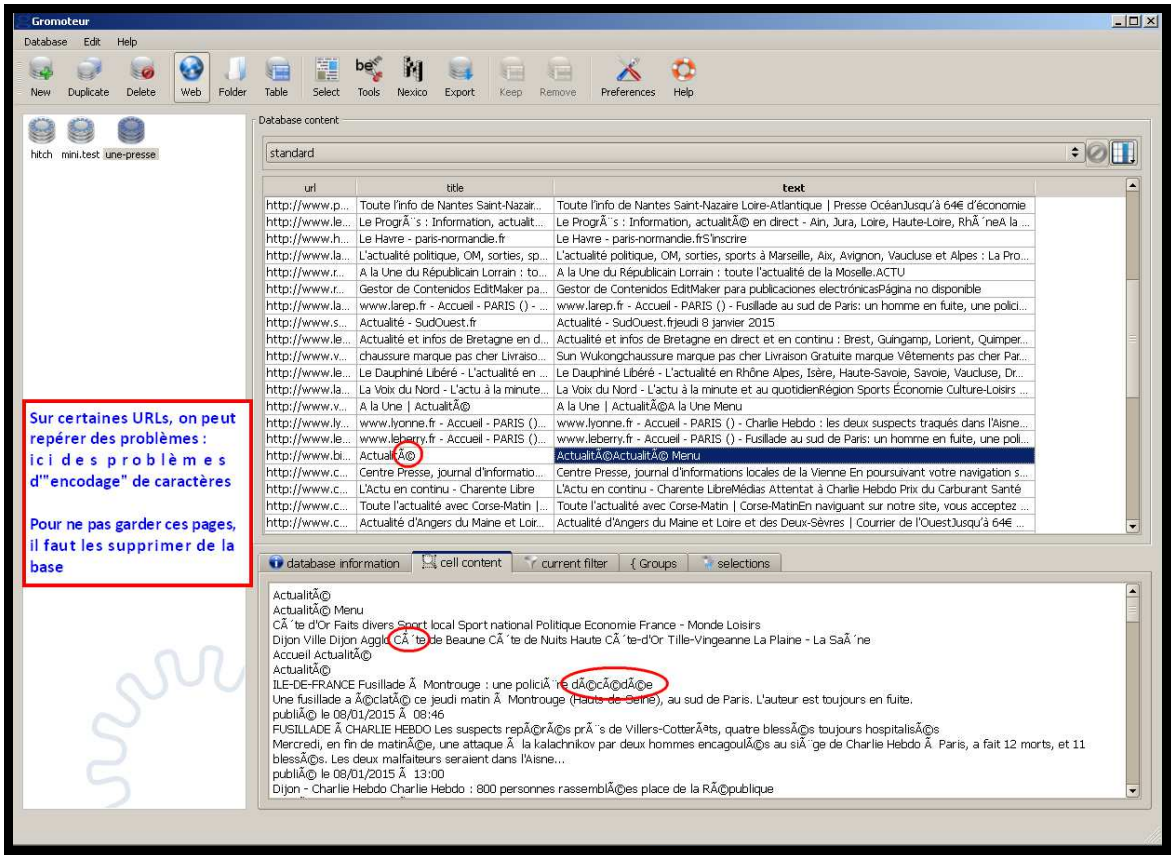
Retour sur la fenêtre principale.

On voit apparaître des informations : le contenu de la base « UNE-PRESSE »

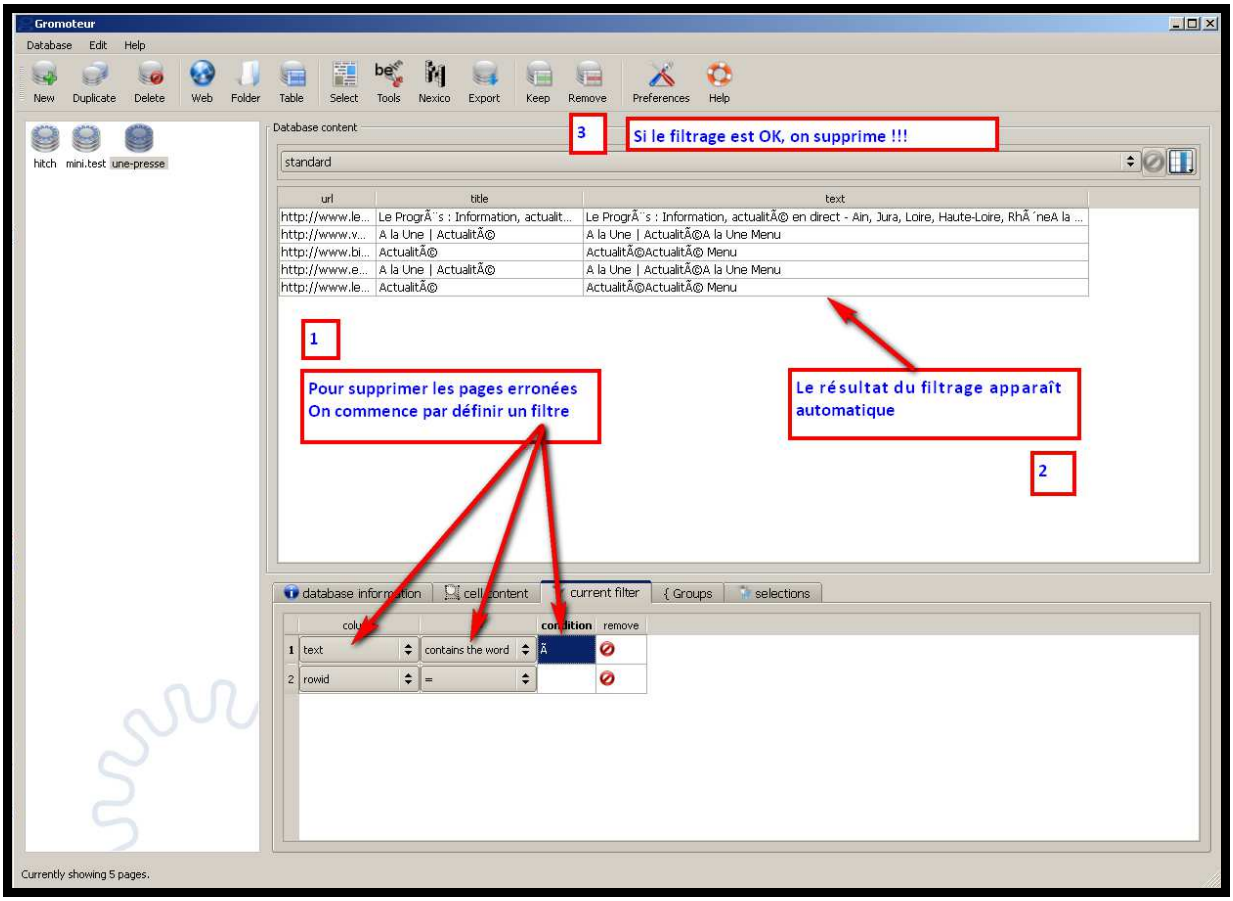


Etape n°8 : Nettoyez le résultat produit par Gromoteur

En examinant les différentes lignes associées chacun à une des URLs de départ, on peut être amené à remarquer des petits problèmes…



Il est préférable dans ce cas, de nettoyer le contenu de la base et de supprimer les lignes visées. Pour cela, on va les identifier en créant un filtre avant des les éliminer de la base.

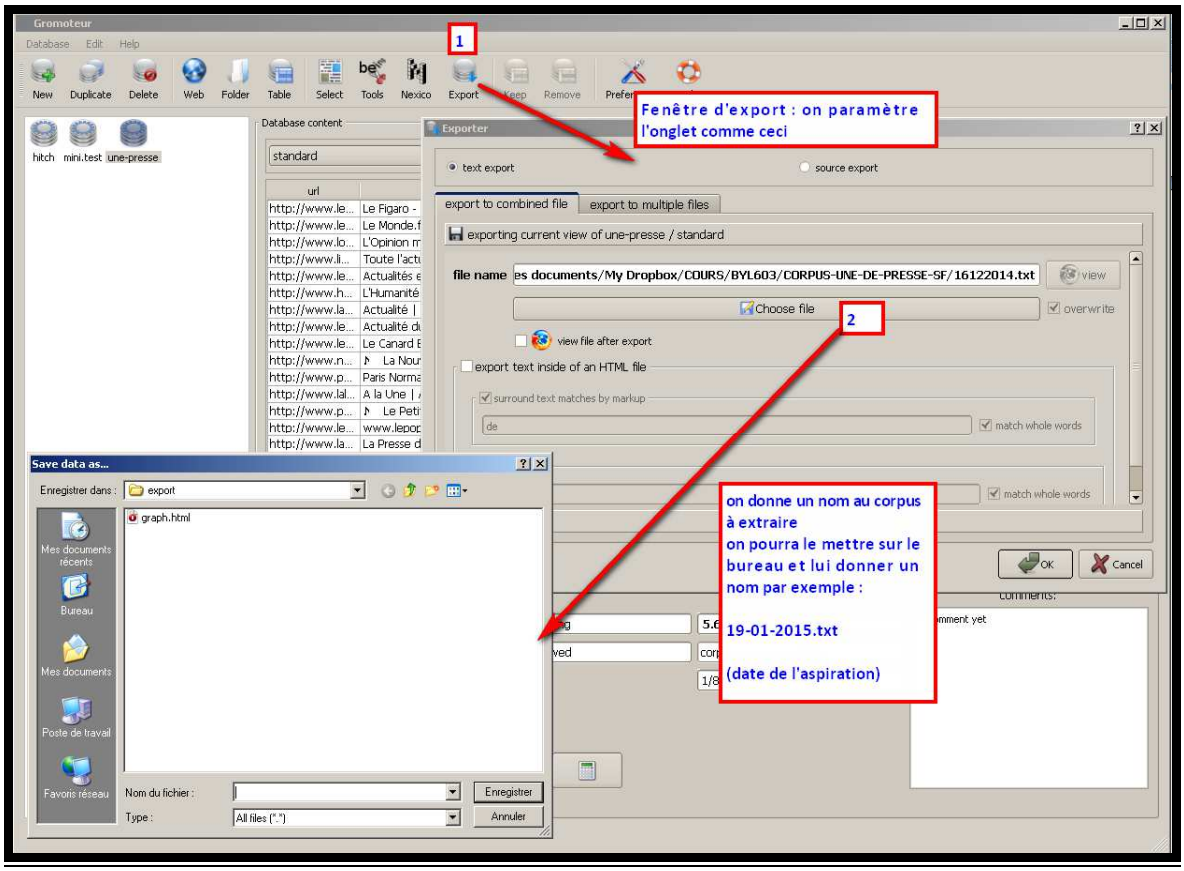


L'opération précédente peut être réitérer autant de fois que nécessaire.

Etape n°9 : Exporter le résultat au format TXT

Une fois la base nettoyée, on va devoir l'exporter.

Il est IMPERATIF de réaliser cette étape pour sauvegarder les données que l'on vient d'aspirer ; en effet, si on relance l'aspiration, la configuration actuellement définie va supprimer le contenu de la base.



Important :

Le fichier exporté est à conserver précieusement.

Il sera réutilisé plus tard.

Pour le projet, si vous conservez ce type de configuration, vous devrez générer un fichier d'export après chaque aspiration et sauvegarder tous les fichiers créés.

Etape n°10 : Construire un corpus chronologique

Les différentes étapes précédentes devront être réitérées un certain nombre de fois pour construire un corpus chronologique (par exemple toutes les semaines pendant la durée du projet.

Tous ces fichiers seront ensuite concaténés pour construire un SEUL fichier qui sera traité par un autre outil (Le Trameur).

Etapes suivantes :

- **Préparer le corpus (FICHE n°2)**

- **Explorer le corpus avec le Trameur (FICHE n° 3)**