

# Fiche n°1 : Construire un corpus avec Gromoteur

---

## Sommaire

Préambule : .....	2
Etape n°1 : Récupérer Gromoteur.....	3
Etape n°2 : Dézipper l'archive téléchargée .....	4
Etape n°3 : Démarrer Gromoteur.....	5
Etape n°4 : Récupérer la liste des URLs à traiter .....	6
Etape n°5 : Paramétrer Gromoteur .....	7
Etape n°6 : Lancer l' « aspiration » .....	14
Etape n°7 : Vérifier le résultat produit par Gromoteur .....	15
Etape n°8 : Nettoyez le résultat produit par Gromoteur .....	16
Etape n°9 : Exporter le résultat au format TXT .....	18
Etape n°10 : Construire un corpus chronologique .....	20
Etales suivantes : .....	21
• préparer le corpus (FICHE n°2) .....	21
• explorer le corpus avec le Trameur (FICHE n° 3) .....	21

## 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

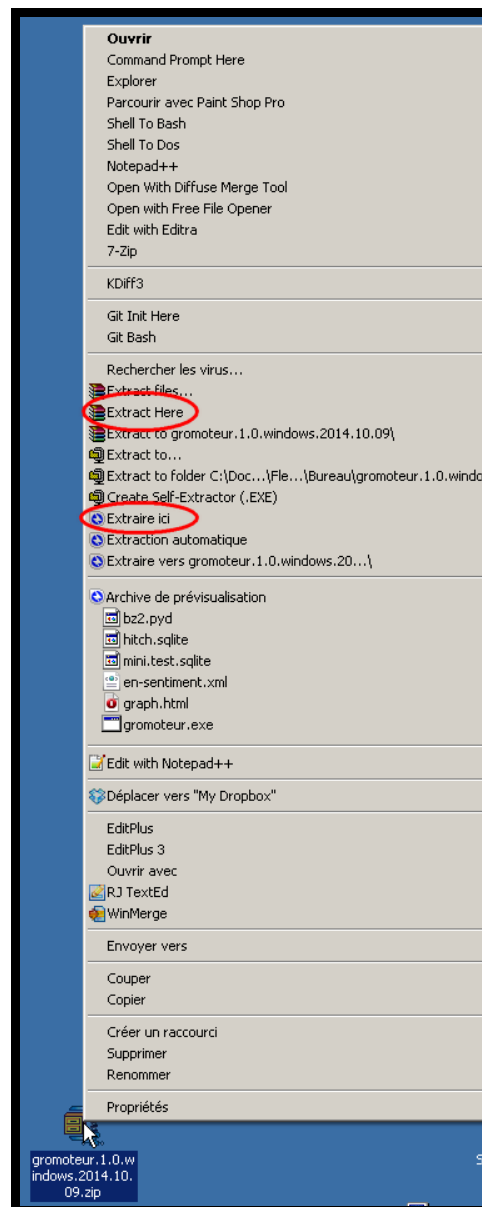
(cf figure ci-dessous)

## Etape n°2 : Dézipper l'archive téléchargée

1. Localisez l'archive téléchargée :

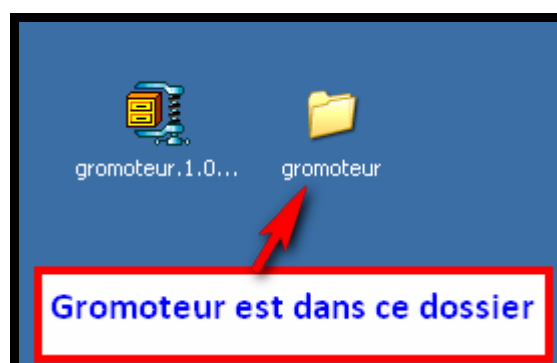


2. Dézippez-la : faire un clic-droit sur l'archive et repérer l'instruction « Extraire ici » ou « extract here » :

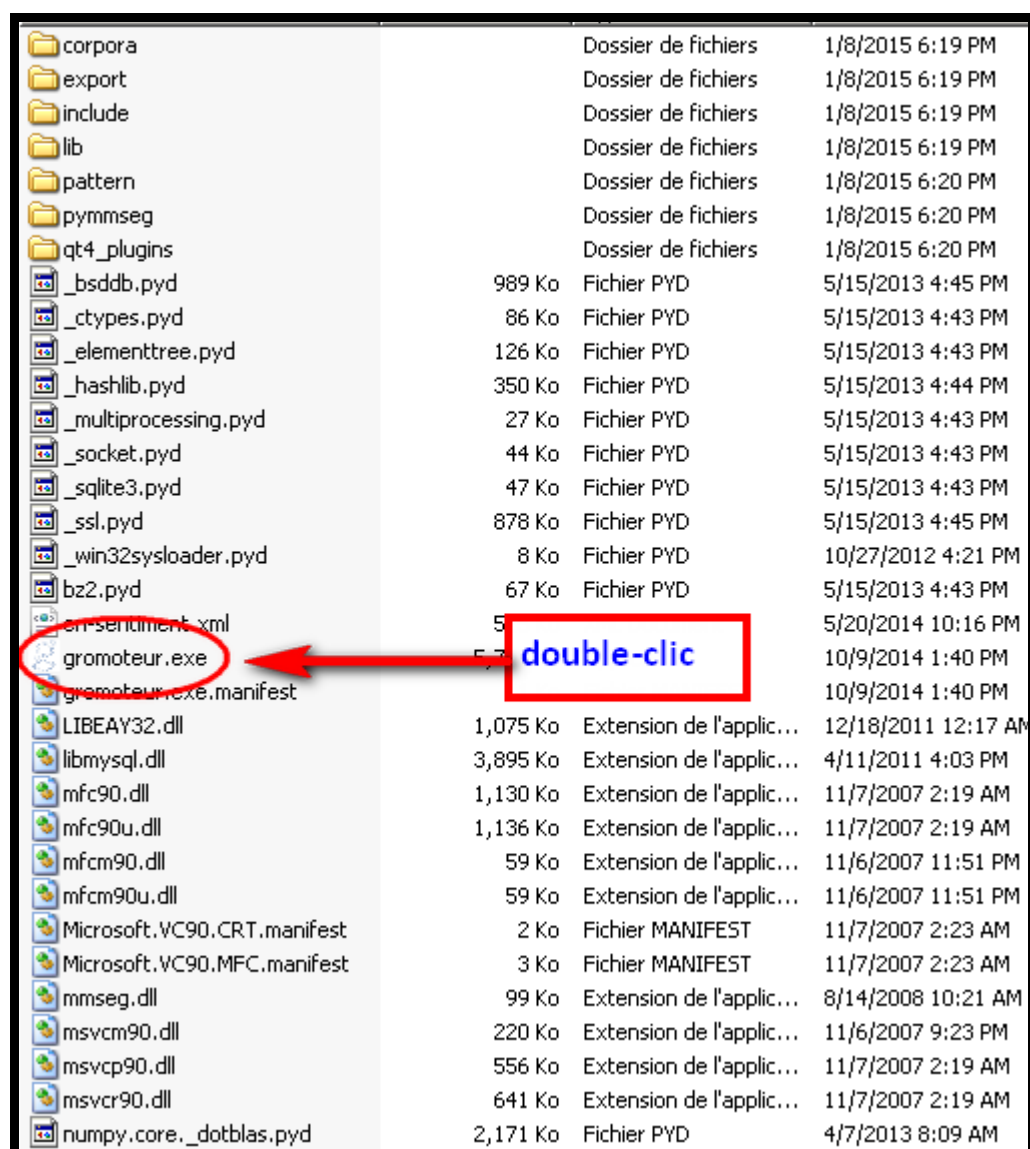


## Etape n°3 : Démarrer Gromoteur

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>

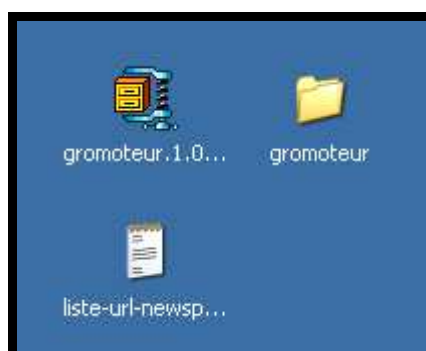
Ce fichier contient des adresses des pages d'accueil de journaux en ligne (nationaux ou régionaux) : il contient 67 adresses.

On en voit ci-dessous un extrait :



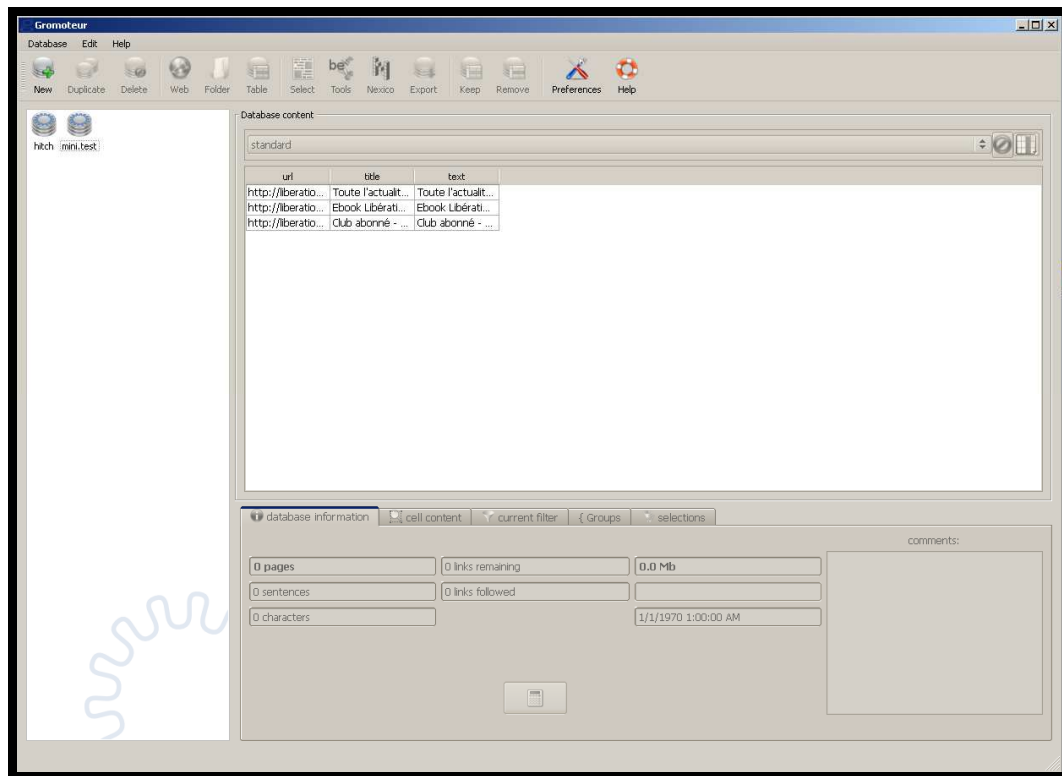
```
liste-url-newspaper.txt
1 http://www.lefigaro.fr
2 http://www.lemonde.fr
3 http://www.lopinion.fr
4 http://www.liberation.fr
5 http://www.leparisien.fr
6 http://www.humanite.fr
7 http://www.la-croix.com
8 http://www.lejdd.fr
9 http://www.lecanardenchaine.fr
```

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

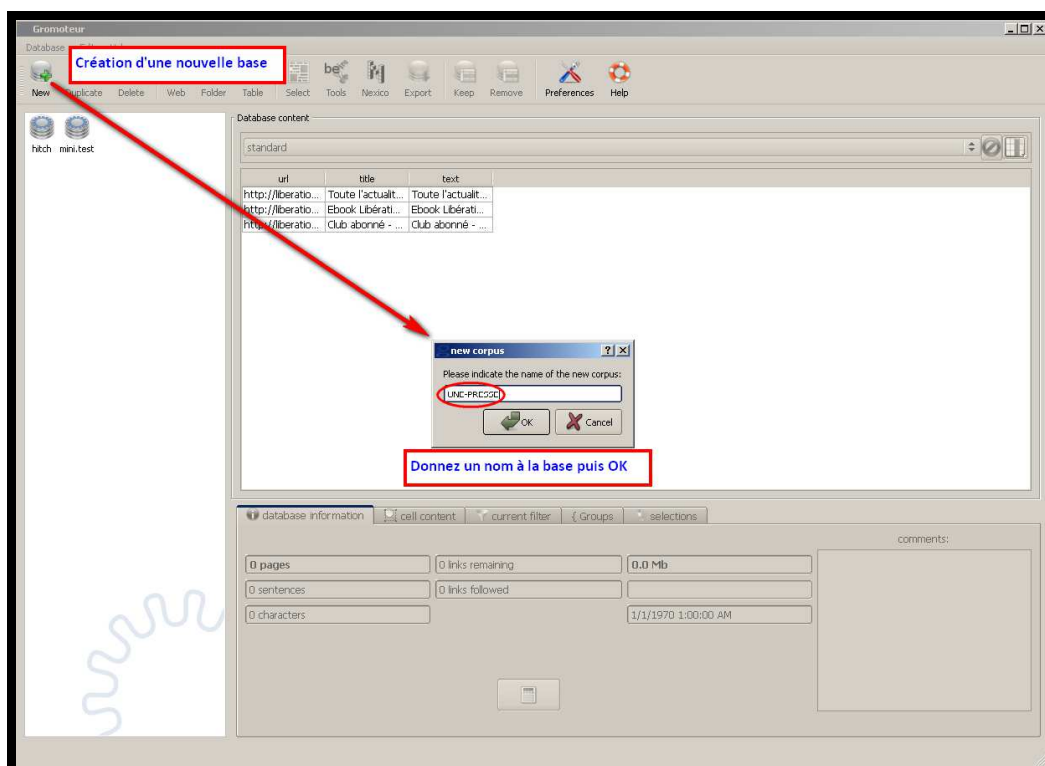


## Etape n°5 : Paramétrer Gromoteur

Une fois lancé, Gromoteur apparaît !



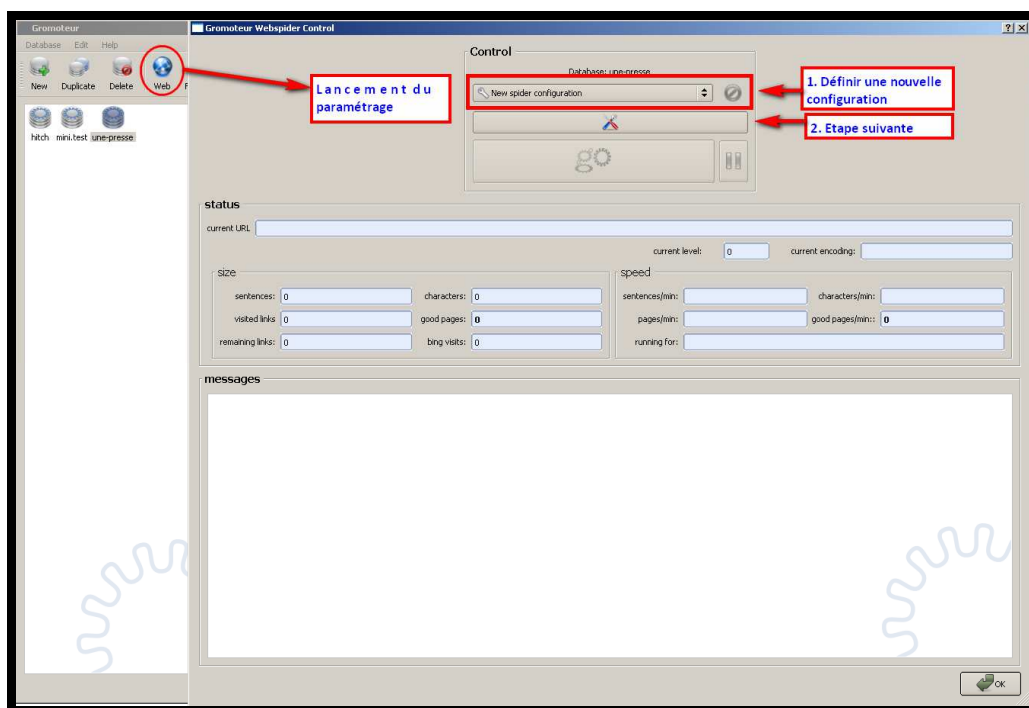
Créer une nouvelle base :



La nouvelle base apparaît : sélectionnez-la

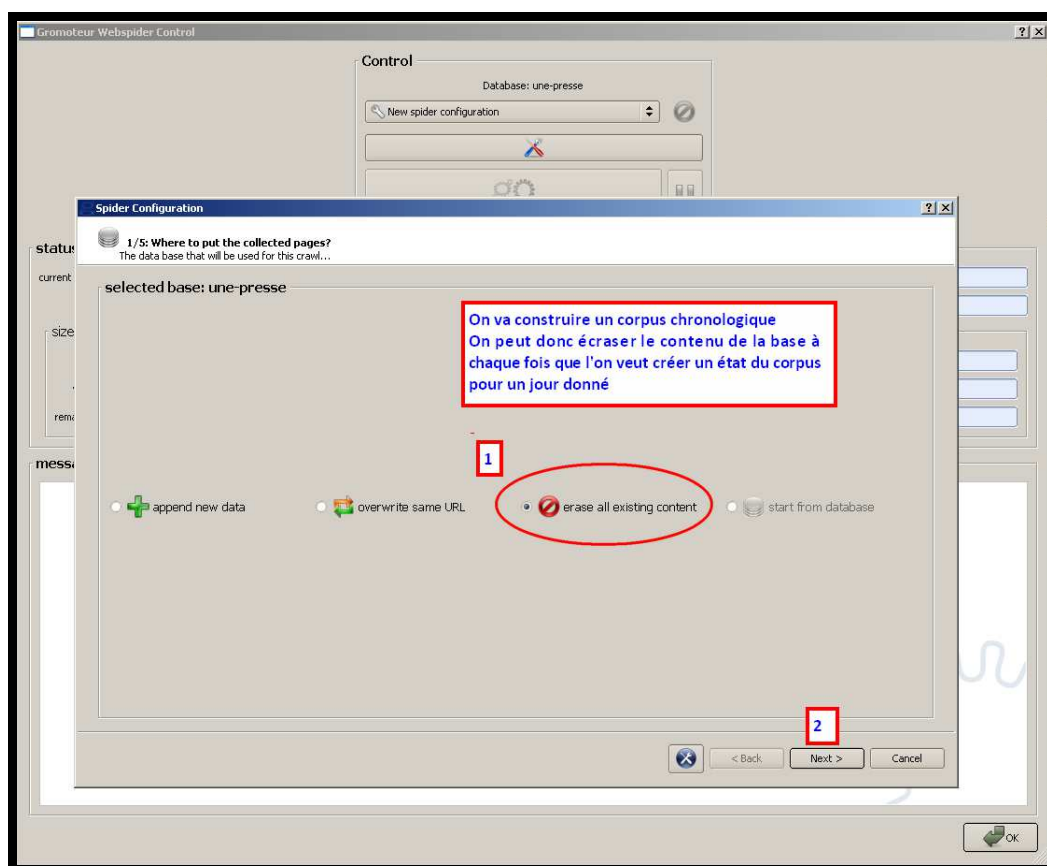


Lancez le paramétrage :



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





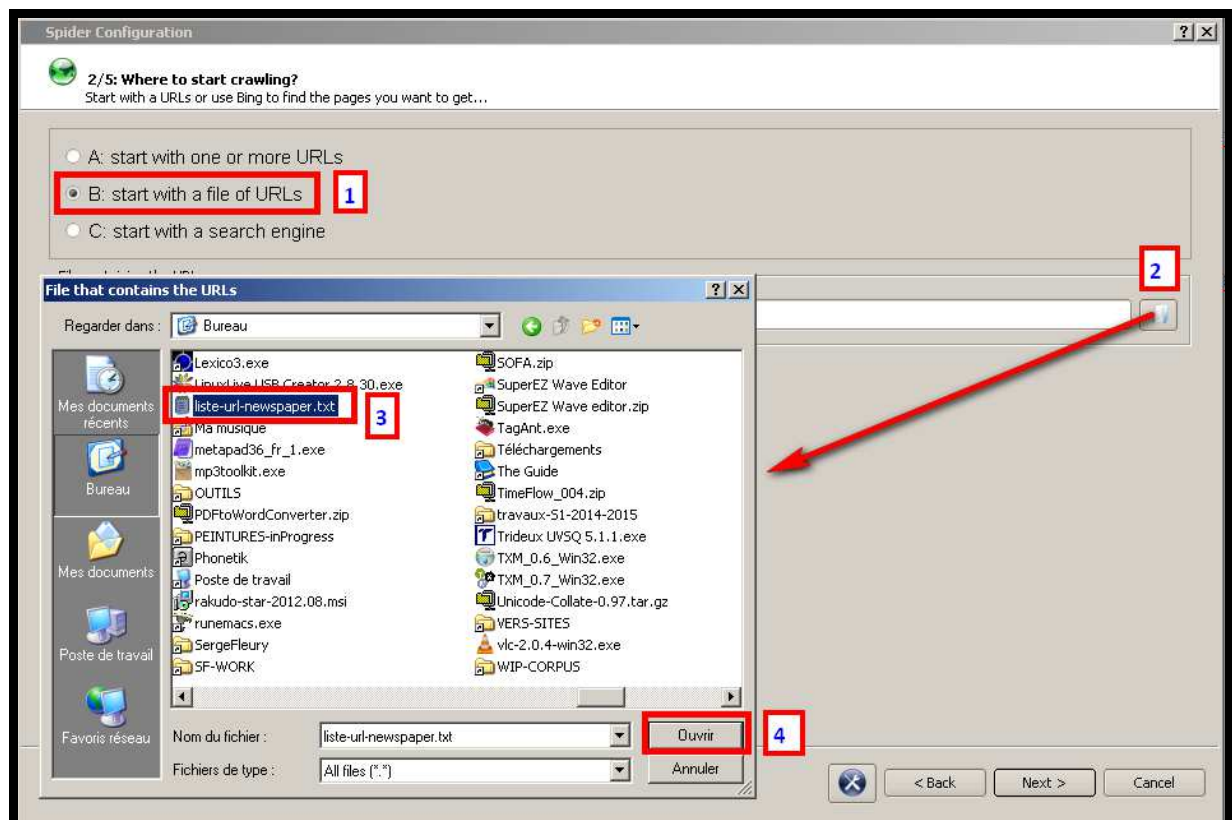
Dans cet onglet, on choisit la manière dont les données seront stockées dans la base.

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.

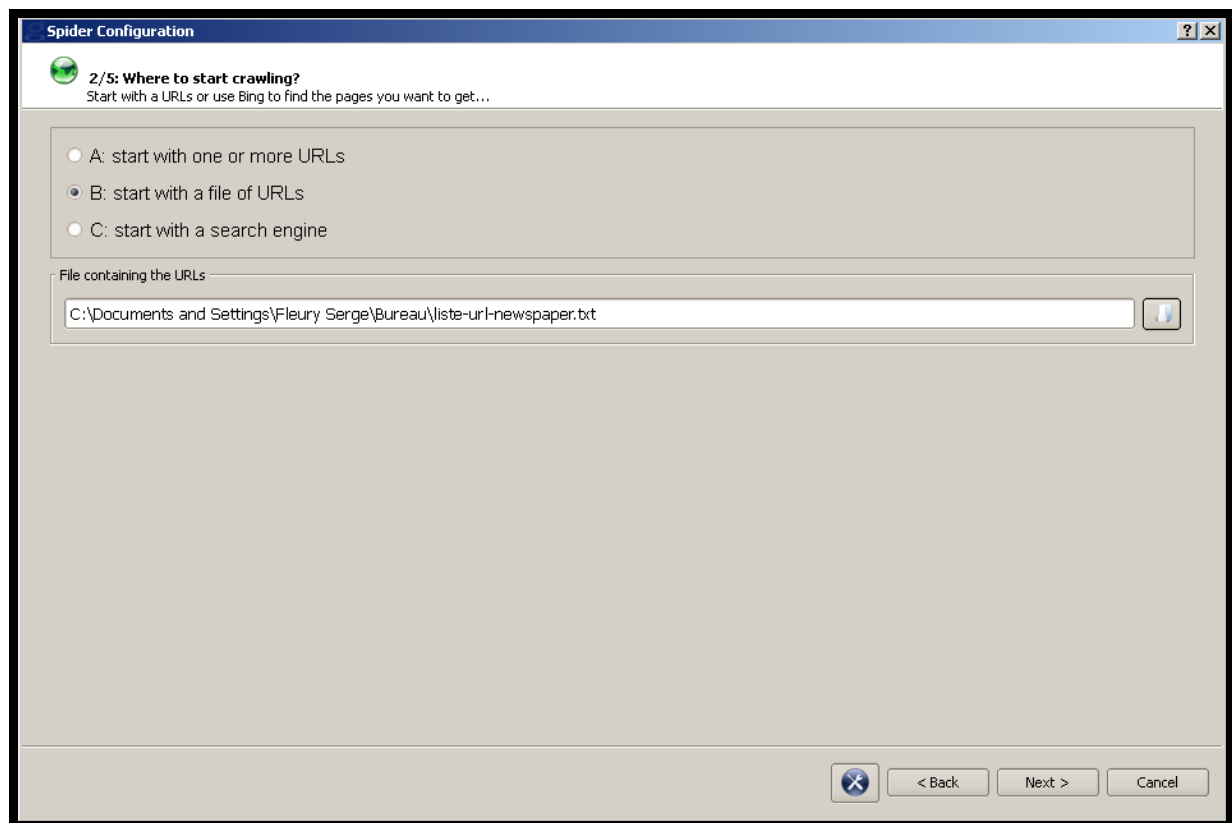
## IMPORTANT :

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.

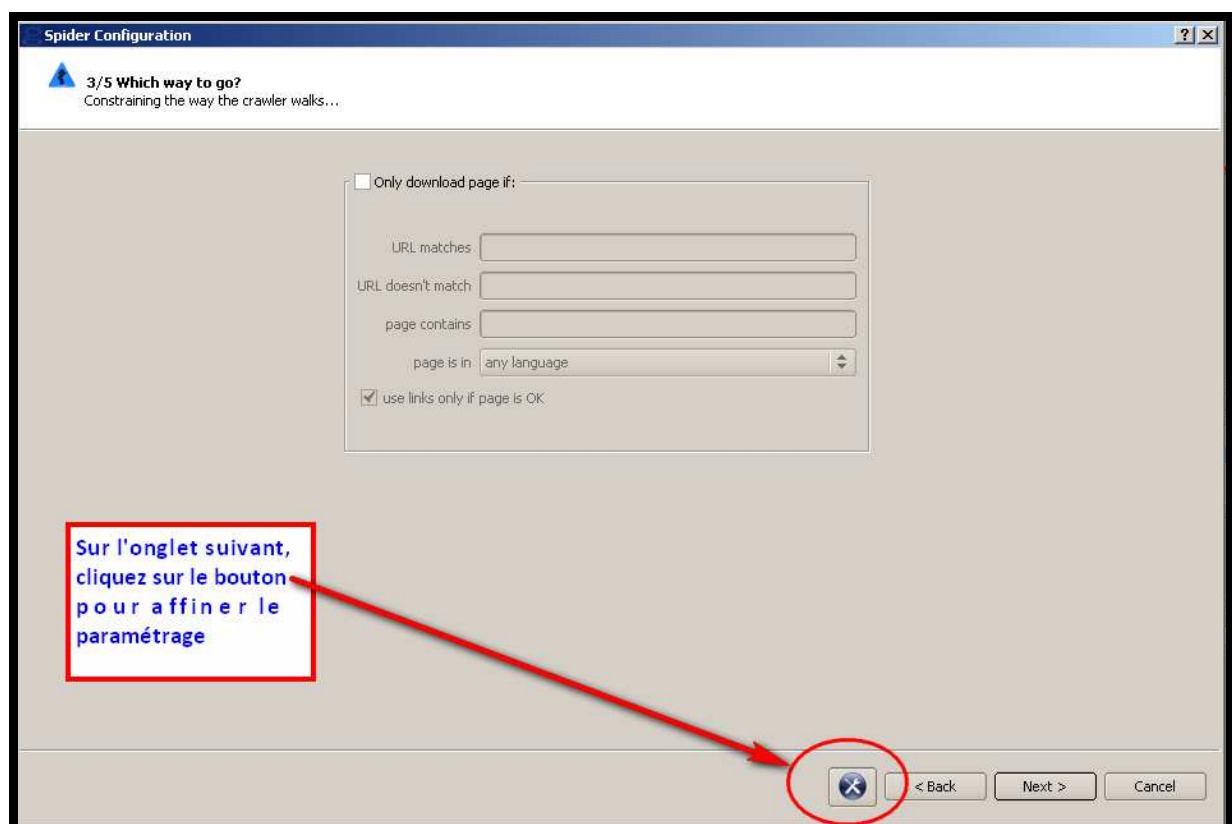


Dans cet onglet, on choisit de paramétrer l'aspiration en sélectionnant un fichier contenant les différentes adresses visées.

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



Après cette étape, tapez sur Next



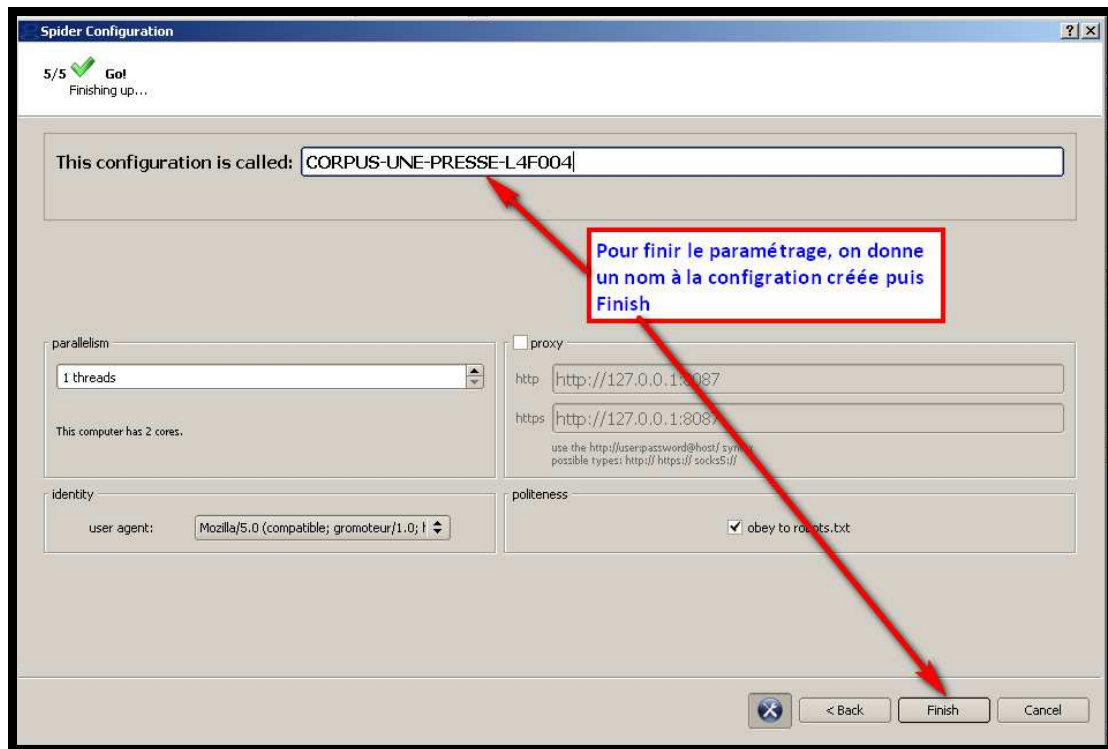
Paramétrez l'onglet suivant comme ceci :

The screenshot shows the 'Spider Configuration' dialog box, step 3/5, titled 'Which way to go?'. The subtitle is 'Constraining the way the crawler walks...'. The 'Path:' section has two radio buttons: 'Breadth first' (selected) and 'Depth first'. Below this, there are two main sections. The left section, 'Only download page if:', is checked and contains fields for 'URL matches', 'URL doesn't match', 'page contains', and 'page is in' (set to 'any language'). It also has 'level from' (0) and 'to' (1) spinners, and checkboxes for 'use links only if page is OK' and 'ignore case'. The right section, 'Only take link if:', is unchecked and contains 'URL matches' and 'URL doesn't match' fields. Below these is a checked 'take pdf files' checkbox and a 'force encoding' field. At the bottom right are buttons for '< Back', 'Next >', and 'Cancel'.

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

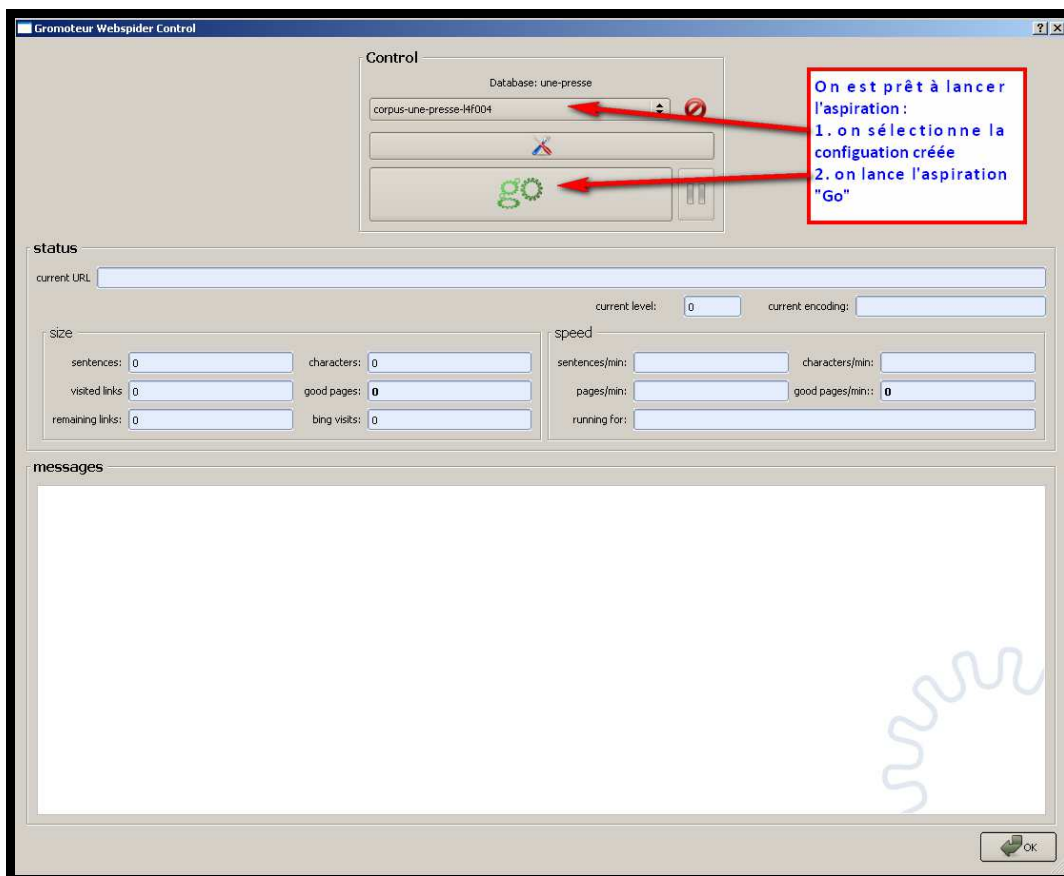
The screenshot shows the 'Spider Configuration' dialog box, step 4/5, titled 'How much?'. The subtitle is 'Constraining the quantity of crawling results...'. It features several input fields: 'max pages' (unlimited), 'max sentences' (unlimited), 'max subdomains' (unlimited), 'max Mb of disk space' (unlimited), and 'keep at least' (0 Mb). A note indicates 'of the 80956.8007812 Mb of free space in the corpus folder'. Below these are two sections. The first, 'Avoid spider traps', is checked and has two options: 'by trying to visit the same server only every' (0 sec) and 'by trying first to take from one server only' (0 pages). The second, 'Follow redirects', is checked and has an unchecked option 'only follow if redirected URL matches download conditions'. At the bottom, the 'Timeout' section is checked with a '1 sec' spinner and a 'try again: 0 times' spinner. At the bottom right are buttons for '< Back', 'Next >', and 'Cancel'.

Après cette étape, tapez sur Next



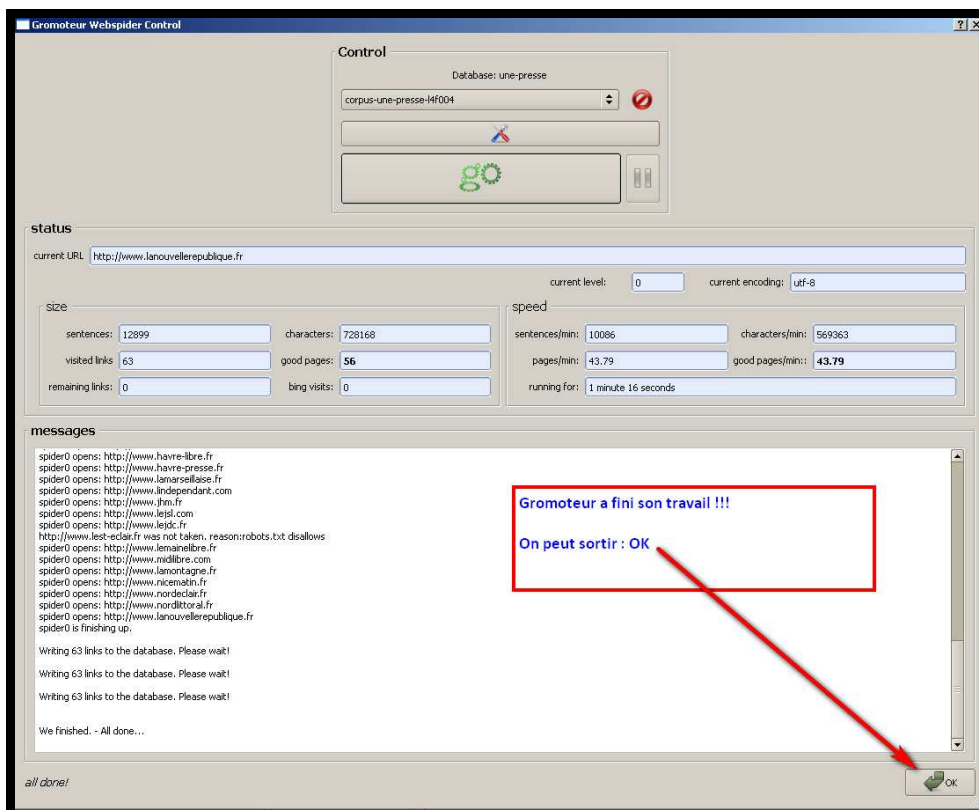
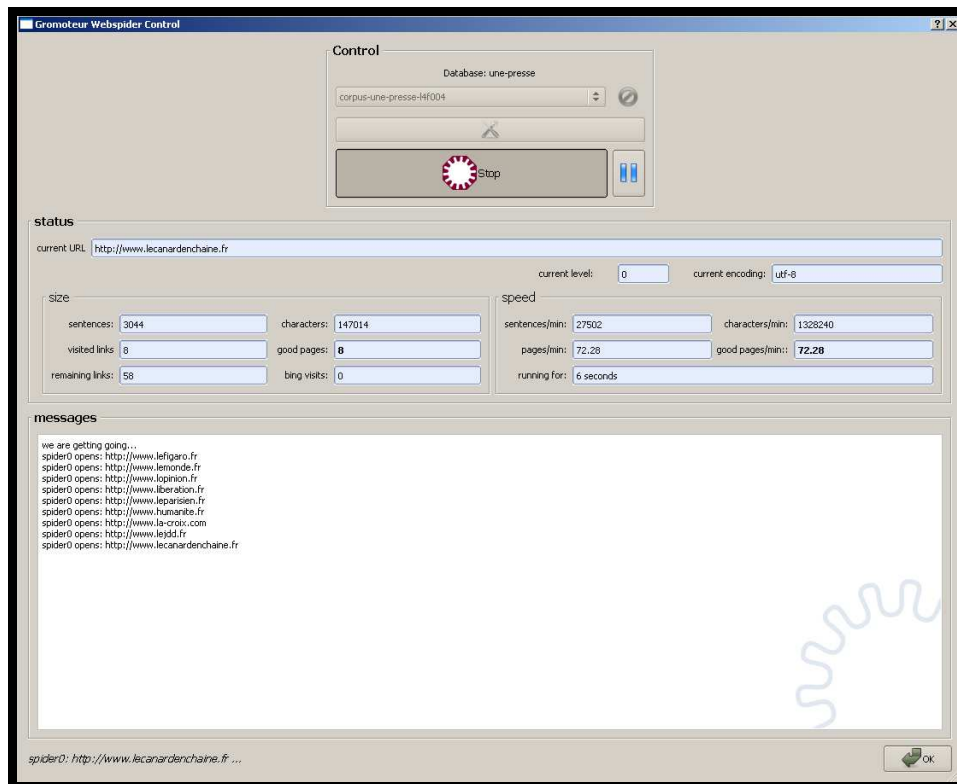
L'onglet précédent est le dernier onglet de paramétrage.

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 »

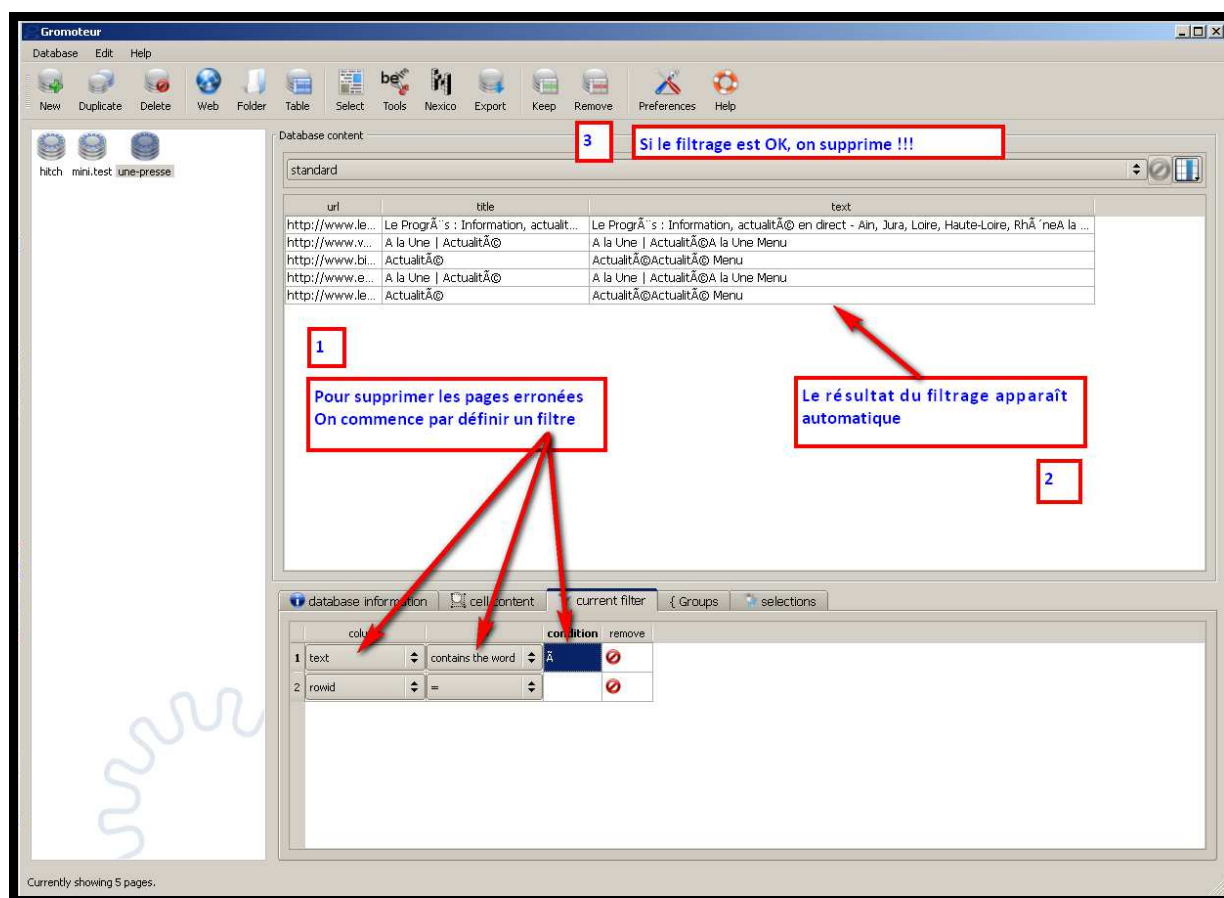
The screenshot shows the Gromoteur application window. On the left, a sidebar lists databases: 'hitch', 'mini.test', and 'une-presse'. A red box highlights a text overlay: 'Gromoteur a rempli la base. Chaque URL est visible sur chaque ligne : le nom de l'URL, son titre et son contenu textuel'. The main area displays a table of database content with columns 'url', 'titre', and 'text'. The table lists various news sources like 'Le Figaro', 'Le Monde.fr', 'L'Opinion média', etc. At the bottom, a 'database information' panel shows statistics: 56 pages, 12899 sentences, 728168 characters, 0 links remaining, 63 links followed, 5.6 Mb, corpus-une-presse-l4f004, and a timestamp of 1/8/2015 6:58:53 PM. A 'comments' section on the right shows 'no comment yet'.

url	titre	text
http://www.le...	Le Figaro - Actualités	Le Figaro - ActualitésLe Scan Politique Economie Sport24 Le Scan Sport Culture Lifestyle M...
http://www.le...	Le Monde.fr - Actualité à la Une	Le Monde.fr - Actualité à la Une En poursuivant votre navigation sur ce site, vous acceptez...
http://www.lo...	L'Opinion média quotidien, libéral, e...	L'Opinion média quotidien, libéral, européen et pro-businessAbonnez-vous Découvrez la Un...
http://www.li...	Toute l'actualité en direct - photos...	Toute l'actualité en direct - photos et vidéos avec Libération - Libération Se connecter S'in...
http://www.le...	Actualités en direct et info en cont...	Actualités en direct et info en continu - Le ParisienEn naviguant sur notre site, vous accep...
http://www.h...	L'Humanité	L'Humanité Jump to navigationL'Humanité.fr , c'est aussi
http://www.la...	Actualité   La-Croix.com	Actualité   La-Croix.comRésultat de la recherche :
http://www.le...	Actualité du jour en direct, politici...	Actualité du jour en direct, politique, culture, médias - leJDD.fr   Se connecter   S'insc...
http://www.le...	Le Canard Enchaîné   Un pavé dans...	Le Canard Enchaîné   Un pavé dans la cyber mare ↓ Skip to Main ContentLes Unes Les Dos...
http://www.n...	La Nouvelle République des	La Nouvelle République des Pyrénées - NRPyrénées.frSe connecter   S'insc...
http://www.p...	Paris Normandie - paris-normandie.fr	Paris Normandie - paris-normandie.frS'inscrire
http://www.la...	A la Une   Actualité	A la Une   ActualitéMenu
http://www.p...	Le Petit Bleu d'Agen - PetitBleu...	Le Petit Bleu d'Agen - PetitBleu.frSe connecter   S'insc...
http://www.le...	www.lepopulaire.fr - Accueil - PARI...	www.lepopulaire.fr - Accueil - PARIS () - Charlie Hebdo : les deux suspects traqués dans l'A...
http://www.la...	La Presse de la Manche - Cherbour...	La Presse de la Manche - Cherbourg Octeville JEUDI 08 JANVIER 2015 CONTACTS QUI SO...
http://www.p...	Toute l'info de Nantes Saint-Nazair...	Toute l'info de Nantes Saint-Nazaire Loire-Atlantique   Presse OcéanJusqu'à 64€ d'économie
http://www.le...	Le Progrès : Information, actualit...	Le Progrès : Information, actualité@ en direct - Ain, Jura, Loire, Haute-Loire, Rhâ...neÀ la ...
http://www.h...	Le Havre - paris-normandie.fr	Le Havre - paris-normandie.frS'inscrire
http://www.la...	L'actualité politique, OM, sorties, sp...	L'actualité politique, OM, sorties, sports à Marseille, Aix, Avignon, Vaucluse et Alpes : La Pro...
http://www.r...	A la Une du Républicain Lorrain : to...	A la Une du Républicain Lorrain : toute l'actualité de la Moselle.ACTU





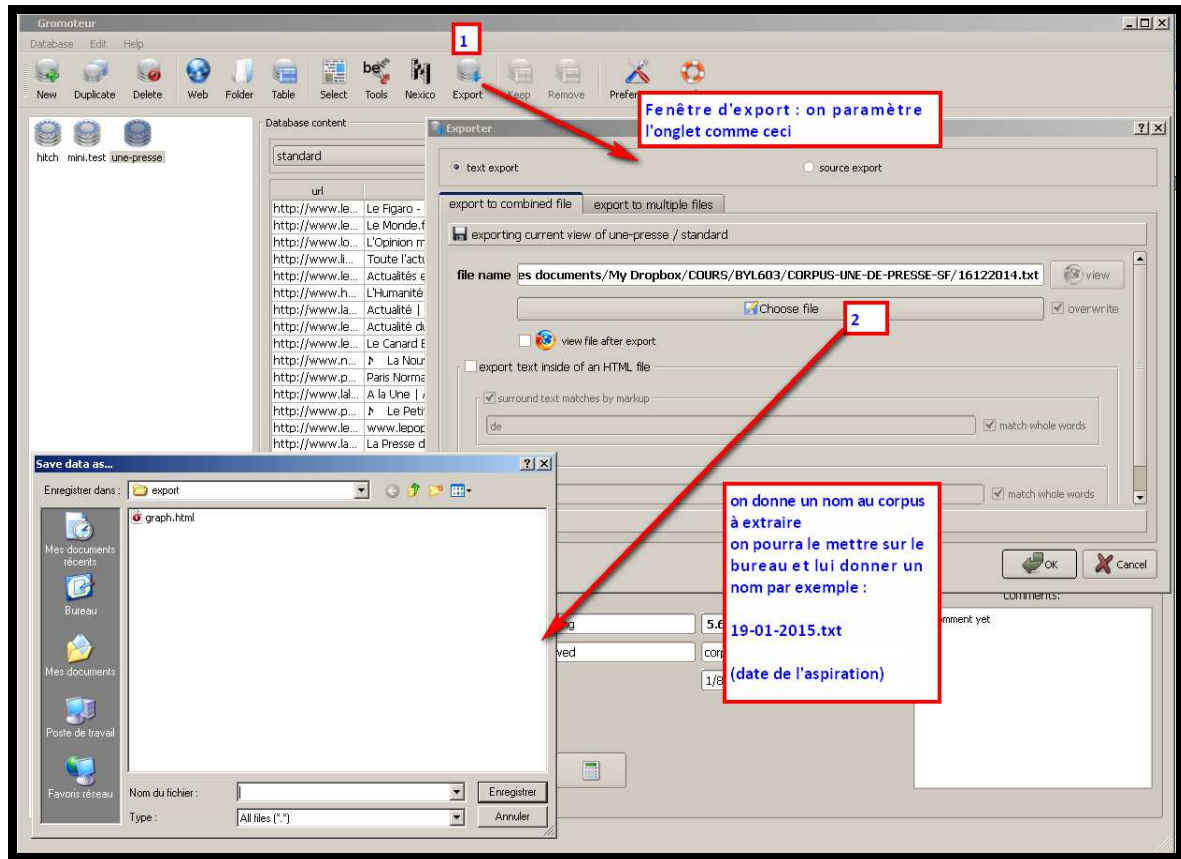
Il est préférable dans ce cas, de nettoyer le contenu de la base et de supprimer les lignes visées. Pour cela, on va les identifier en créant un filtre avant de les éliminer de la base.



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

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 répétées 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)**