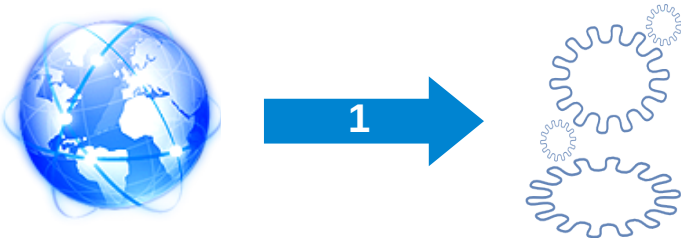
Gro&Rapide

Ce document explique simplement comment utiliser Gromoteur pour

- télécharger un site
- nettoyer le texte
- exporter le texte

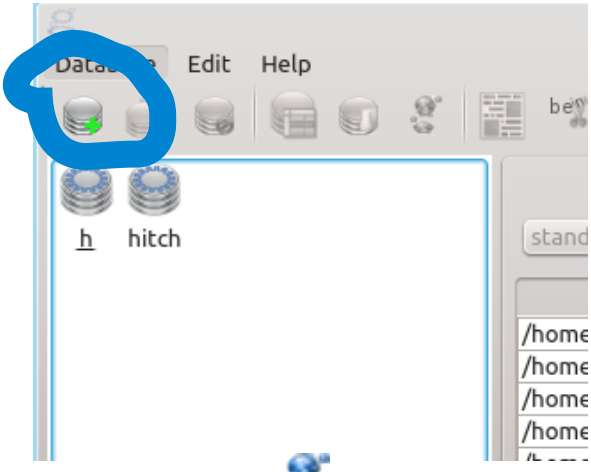
> - Noter que pratiquement tous les champs de Gromoteur ont une infobulle (tooltip) expliquant son utilisation : mettez la souris dessus et attendez !
> - Il est très utile d'avoir une certaine maîtrise d'expression régulières pour profiter de tous les choix de configuration du Gromoteur. Vous pouvez pratiquer ici : http://elizia.net/regex/

1 Du web vers une base de Gromoteur



Recueillir des pages web vers une base de Gromoteur en 10 étapes :

1. Créer une nouvelle base. Donner un nom au choix. OK. La base apparaît dans le panneau de bases.

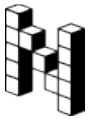


2. Cliquer sur le bouton « g »  pour ouvrir le robot d'indexation (spider). La fenêtre du contrôle du spider s'ouvre.
3. Cliquer sur le menu déroulant 

et sélectionner « New spider configuration » (Ou, si vous avez déjà une configuration, la choisir)
4. Cliquer sur le bouton  . La fenêtre de la configuration du spider s'ouvre.

5. Le bouton  permet de choisir entre mode normal et mode expert :
   - mode simple : les choix essentiels
   - mode expert : toutes les configurations du spider de Gromoteur
6. La configuration a 5 étapes :
   1. Où mettre les pages sélectionnées ?
   2. Où commencer à chercher ?
   3. Où aller ?
   4. Combien prendre ?
   5. Comment appeler la configuration ?

   Les étapes sont expliquées ci-dessous
7. À la fin, on clique sur « Finish » pour retourner à la fenêtre du contrôle du spider.
8. Maintenant, on clique sur le bouton de démarrage  pour démarrer la collection de pages.
9. La collecte de pages peut s'arrêter pour différentes raisons :
   - le nombre maximum de pages a été atteint ou le disque est trop plein
   - il n'y a plus de liens disponibles pour continuer la collecte
   - l'utilisateur a cliqué sur  pour arrêter la collecte. Dans ce dernier cas, il faut attendre que chaque thread ait terminé son travail et le bouton de démarrage redevienne vert.

   On ferme la fenêtre de contrôle du spider en cliquant sur « OK ». On accède ainsi à nouveau à la fenêtre principale de Gromoteur.
10. Le tableau principale montre maintenant toutes les pages collectionnées, une par ligne. On peut maintenant
    - restreindre la vue de tableau à l'aide de filtres 
    - exporter le code source des pages dans un ou plusieurs fichiers 
    - exporter le texte brute dans un ou plusieurs fichiers (voir la section « Export » ci-dessous)
    - extraire systématiquement certaines parties de la page  dans une nouvelle colonne du tableau, par exemple afin de disposer du texte simple de la page web sans publicité etc. autour. Voir section spécifiques ci-dessous.
    - grouper les pages afin de les exporter ou analyser statistiquement conjointement. {
    - analyser les pages statistiquement à l'aide de Nexico! 

2 La configuration du spider pour la collecte de pages Web

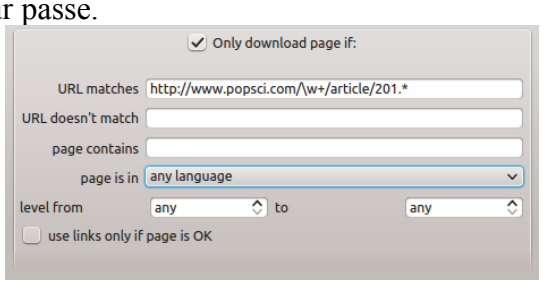
Les règles de base de la configuration :

- procéder par peu de contraintes, tester avec qqs pages et ajouter des contraintes jusqu'à ce qu'on obtient des pages souhaitées. Seulement en ce moment, on passe à une plus grande quantité de pages.
- En cas de doute, laisser les configurations par défaut.

Les cinq étapes ont plus de possibilités de configuration quand on passe en mode expert en enfonçant le bouton  .

1. **Où mettre les pages sélectionnées ?**
   4 possibilités :
   1. *append new data* : ajouter les nouvelles pages à la fin du tableau existant, indépendant d'une éventuelle identité des urls.
   2. *overwrite same URL* : si l'URL est identique, écraser l'ancienne page.
   3. *erase all existing content* : effacer la base (si elle contient déjà qqch)
   4. *start from database* : prendre les URL qui restent à faire pour continuer la collecte de pages (seulement accessible si la base contient déjà des liens pour continuer).

   En cas de doute, choisissez *3. erase all existing content*
2. **Où commencer à chercher ?**
   3 possibilités :
   1. commencer avec un ou plusieurs URL (plusieurs URL peuvent être séparés par des espaces) : mettez le ou les URL dans la ligne des URL
   2. fournir un fichier contenant les URL du départ, un URL par ligne
   3. commencer par une requête de moteur de recherche. Gromoteur utilise Bing.
      - Si vous avez beaucoup de requêtes à faire (plus de 10 par jour), obtenez votre propre compte sur www.bing.com/developers/createapp.aspx afin de ne pas gêner d'autres utilisateurs de Gromoteur.
      - Mettre les mots clés. Pour voir les options de requête : http://onlinehelp.microsoft.com/en-us/bing/ff808421.aspx
      - Tester si la requête marche avec les boutons « try Bing » et « Firefox »
      - contraindre la collecte de résultats en donnant une expression régulière sur les URL à prendre (dans la case « follow only »)
3. **Où aller ?**
   Si vous êtes *en mode normal*, vous n'avez qu'à contraindre les URL par lesquels vous voulez que Gromoteur passe.



Si cette boîte est cochée, le Gromoteur comparera chaque page avec les conditions données avant de l'ajouter à la base. On peut contraindre la correspondance avec un URL (dans l'image ci-dessus, on voit qu'on ne veut que des pages du site « popsci.com » dans l'URL contient une suite de lettre (\w+), le mot « article » et une date commençant par 201. On peut donc donner des contraintes positives et négatives sur l'URL, sur le contenu texte de la page et sur la langue de la page (test basé sur des trigrammes, voir

http://www.elizia.net/languageDetector/ pour plus d'informations). En plus, on peut contraindre le niveau auquel on souhaite récupérer : le niveau 0 correspond aux URL donnés ou obtenus en étape 2 (où commencer à chercher ?), le niveau 1 sont les liens trouvés sur ces pages, etc. (La distinction de niveaux n'est pas tout à fait arborescente, car il peut y avoir des liens sur les pages plus basses dans la liste qui pointe vers les niveaux supérieurs.). Finalement, on peut cocher « use links only if page is OK »

- Si coché : Seulement si la page remplit les conditions données, les liens sont extraits et ajoutés à la liste des liens à visiter dans la suite
- Si décoché : Les liens sont extraits et ajoutés à la liste des liens à visiter dans la suite même si la page elle-même ne remplit pas les conditions données.

En *mode expert*, on a quelques choix de plus :

- Le chemin : soit visiter les liens en ordre de niveau : d'abord niveau 0, ensuite niveau 1 etc (breadth first), soit plonger d'abord en profondeur (depth first) en allons le plus profondement en suivant toujours le premier lien (fourni ou trouvé).
- On peut fournir des conditions séparées quand un lien est ajouté à la liste des sites à visiter.
- On peut inclure des fichiers pdf à la collecte.
- On peut fournir l'encodage des pages à récupérer si on constate que le Gromoteur se trompe.

4. **Combien prendre ?**

   On donne le nombre de pages qu'on souhaite récupérer. Il est utile de commencer avec un petit nombre (10 à 30) pour commencer. On peut aussi fournir ce chiffre en nombre de phrase (estimé par une simple segmentation basée sur la ponctuation). Alternativement, on peut limiter la collecte par la taille de la base de données ou par l'espace restant sur le disque dur utilisé pour stocké la base de données.

   En *mode expert*, on peut aussi restreindre les sous-domaine (xxxx.domaine.com), utile pour la récupération de sites de blogs etc.

   Cette page de configuration propose aussi d'autres paramétrage :

   - éviter des pièges à spider avec des heuristiques simples : contraindre la vitesse avec laquelle le Gromoteur visite chaque serveur et en essayant de changer l'ordre entre les liens à récupérer de manière à ce que chaque serveur est sollicité peu fréquemment.
   - Suivre ou non les redirections d'URL (l'URL change quand on va sur un lien)
   - Il faut abandonner après combien de temps et il faut réessayer de récupérer la page du serveur combien de temps ?

5. **Comment appeler la configuration ?**

   En *mode normal*, on n'a qu'à nommer la configuration qu'on vient de faire.

   En *mode expert*, on décide encore

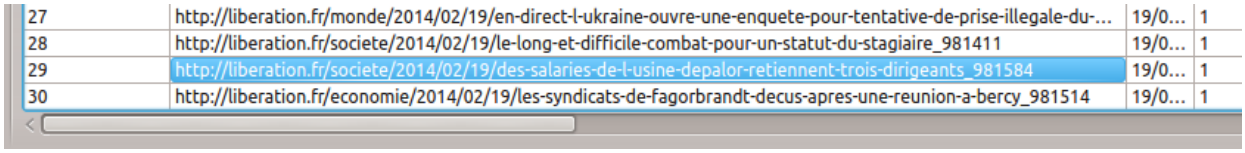
   - combien de fils d'exécution seront lancés en parallèle. Le bon nombre dépend du nombre de cœur de l'ordinateur, la vitesse de la connection internet et la vitesse du serveur visité. Il faut faire des essais.
   - Si on passe ou non par un proxy (http, https, socks5).
   - Si on se présente sous le nom de Gromoteur ou sous une autre identité.
   - Si on obéit ou non au fichier *robots.txt*. Voir http://en.wikipedia.org/wiki/Robots_exclusion_standard pour plus d'information.

On ferme la configuration en cliquant sur « Finish » et on lance la collecte avec le bouton  .

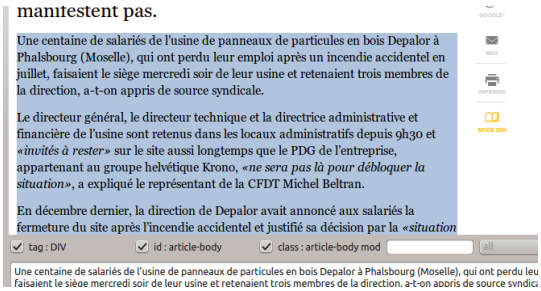
3 Nettoyage des pages Web collectées

Beaucoup de pages Web contiennent des parties qui ne sont pas intéressantes pour l'exploration linguistiques : la partie générique de chaque page (Titre du journal, de la section, liens vers d'autres sections, mentions legales) et les publicités. Souvent les parties qu'on veut retenir sont marquées par des balises html spécifiques. Le Gromoteur permet de découvrir ces parties et d'extraire que les parties intéressantes :

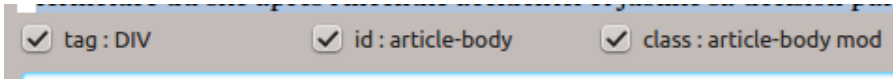
1. dans la fenêtre principale, ouvrir la base à traiter, sélectionner une page qui contient des parties qu'on souhaite extraire :



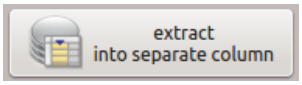
2. Cliquer sur  pour ouvrir le sélecteur de champs.
3. Sélectionner une des parties de la page à extraire en cliquant dessus. Elle se colorise en bleu :



4. On voit en bas de la page sur quels critères la partie a été choisie :



5. Tester si les critères tiennent la route sur d'autres pages en cliquant sur « previous » et « next »
6. Si c'est bon, choisir une textualisation (*standard* par défaut) et donner un nom à la colonne du tableau qui contiendra la partie extraite :

7. cliquer sur le bouton 

pour lancer l'extraction de la partie choisie de toutes les pages actuelles du tableau.
8. Si on n'a pas d'autres parties des pages à extraire, on peut maintenant fermer la fenêtre et retourner dans la fenêtre principale du Gromoteur. On y trouvera la (ou les) nouvelles collones extraites à l'extrême droite du tableau.

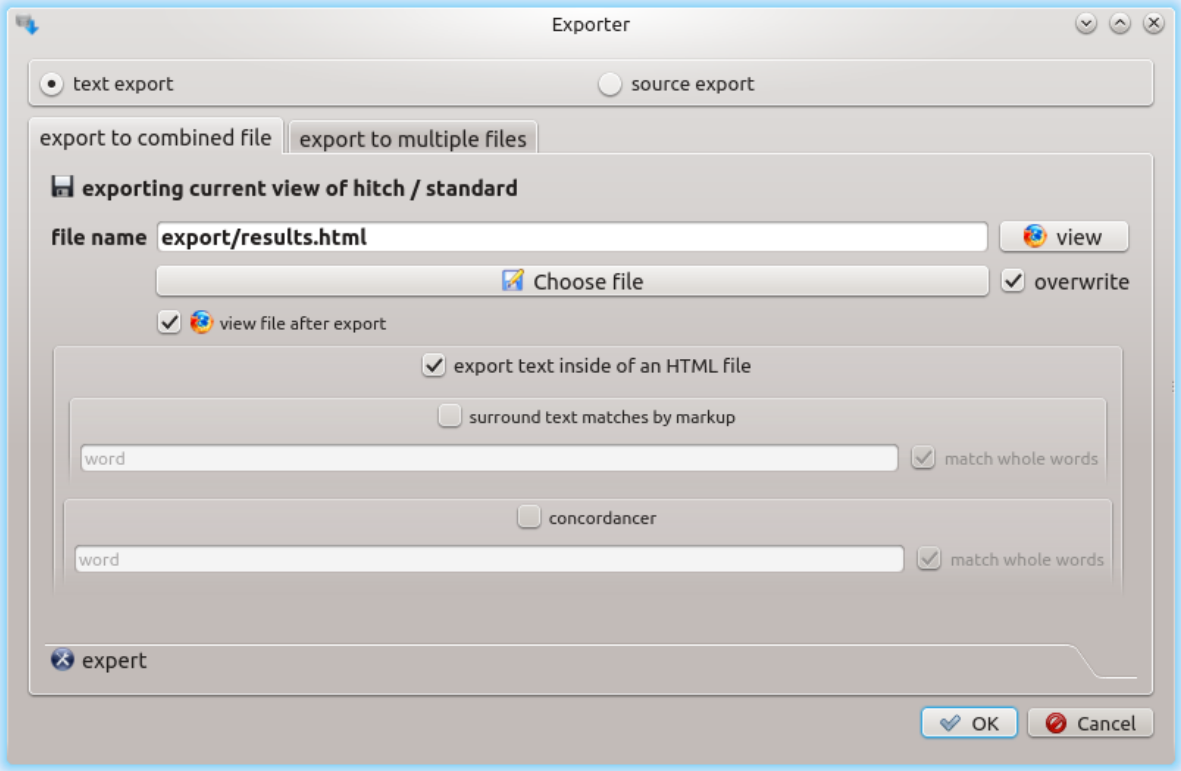
4 L'export

Gromoteur permet d'exporter le contenu d'une base de donnée dans des fichiers texte.

À la base, on a le choix entre l'export du code source HTML ou du texte.

Pour exporter du texte, on procède comme suit :

1. Sélection du contenu à exporter dans la fenêtre principale :
   a) Choix des colonnes : Clique-droite sur l'en-tête de la table principale permet de cacher ou de montrer le contenu. Le menu  permet aussi de choisir les colonne souhaités.
   b) Choix de lignes : Grâce à l'onglet "current filter"  on sélectionne les lignes souhaitées.
2. Le bouton  (ou le menu *Database*, puis *Export*, ou encore le raccourci CTRL-E) ouvre l'exportateur, qui permet d'exporter le contenu de la base dans des fichiers de types différents :



3. Le choix par défaut est celui où le contenu textuel est envoyé dans un fichier html unique, qui est créé et qui s'ouvre quand on clique sur OK. L'exportateur utilise un modèle (template) qu'on peut choisir dans les options « expert »  en bas de la fenêtre.
4. On peut se faire colorer un mot en le mettant à la place de « word » dans la ligne sous « surround text matches by markup » après avoir coché l'option.

5. On peut également se faire exporter les concordances d'un mot en le mettant à la place de « word » dans la ligne sous « concordancer » après avoir coché l'option. Cela donne un fichier html qui ressemble à ceci :

20.	To Send People On A One-Way Trip To Mars 73418 359 7 Why We Can't Stop Eating Frosting From The	Can	73348 359 8 Wing And A Scare 72205 359 9 Dear Congress: Why Are You So Anti-Science? 73263 359 10
http://www.popsci.com/science/article/2013-05/how-avoid-meeting-neanderthals-fate			
21.	What Modern Humans	Can	Learn From The Neanderthals' Extinction | Popular Science
22.	What Modern Humans	Can	Learn From The Neanderthals' Extinction | Popular Science Login/Register Newsletter Subscribe
23.	science 73784 What Modern Humans	Can	Learn From The Neanderthals' Extinction It's a fact of the archaeological record: Modern humans
24.	bones, tools, and pieces of art—along with some DNA that modern humans inherited from them. How	can	we avoid meeting the Neanderthals' fate? That depends on what you think wiped out these early
25.	Extermination and Assimilation The complicated debate over what happened to Neanderthals	can	be boiled down to two dominant theories: Either H. sapiens destroyed the other humans, or joined
26.	ancestry back to a single H. sapiens woman from Africa, nicknamed Mitochondrial Eve. If all of us	can	trace our roots back to one African woman, then how could we be the products of crossbreeding? We
27.	More evidence for Hawks's claims comes from Neanderthal DNA. Samples of their genetic material	can	reveal just what happened after all that Pleistocene hanky-panky. A group of geneticists at the
28.	Several of those regions contain genes connected to the neurological connections that humans	can	form in their brains. In other words, it's possible that H. sapiens ' greater capacity for
29.	many times over. And it spawned deadly famines, too. Humanity's old community-building habits	can	become pathological on a mass scale. Thousands of years after the merging of Neanderthals and H.
30.	Group (Canada), a division of Pearson Canada, Inc. Previous Article: Electrical Brain Stimulation	Can	Help You Learn Math Next Article: FYI: Which Emotion Is The Hardest To Fake? 16 Comments Link to
31.	and battles. They were hunter gatherers, not the civilization builders of the later B.Cs. Why	can	't PopSci writers actually research what they write about? Your dedicated readers are people who
32.	that. Link to this comment mike13323 05/16/13 at 6:58 pm If you want me to go into details John I	can	provide you with evidence supporting all of my criticisms. First, recent carbon dating has moved
33.	40,000 years ago. Only with an open mind and a willingness to look at all angles of the equation,	can	we hope to GUESS at what our ancestors thought. Ideas and beliefs are such an ethereal thing that
34.	to this comment GodLikesComedy 05/17/13 at 12:13 pm Sorry I didn't read everything, but how we	can	survive? 1 STOP WARS 2 STOP FIAT CURRENCIES 3 STOP BORDERS 4 STOP CREATING WEAPONS 5 ST
35.	and theorize about the past is quite the debate in archaeology and paleoanthropology. The best we	can	do at the moment is try to not corrupt the past with modern notions like we've mentioned above.
36.	Of Darker Skin Makes Them Less Racist Space Tourism's Black Carbon Problem What Modern Humans	Can	Learn From The Neanderthals' Extinction Untouched For The Last Billion Years, Water In Canadian
37.	Climate Modeling Method Wanna Know How You're Going To Survive The Apocalypse? This Bacterium	Can	Do Division, Compute Logarithms And Take Square Roots Most Viewed Science The Week In Numbers:

6. Pour les deux choix précédents, l'utilisation d'expressions régulières est possbible.
7. Le 2e onglet « export to multiple files » permet d'obtenir un fichier par ligne du tableau. Les noms du fichiers peuvent être les URLs de la ligne ou un simple numéro.

Pour exporter le code source, on procède comme suit :

1. On choisit les lignes qu'on souhaite exporter, la sélection de colonnes visibles n'entre pas en jeu.
2. On lance l'exportateur.
3. On choisit « source export » en haut de la fenêtre.
4. On choisit entre fichier unique et fichiers séparés en ouvrant l'onglet correspondant.
5. On clique sur OK.