

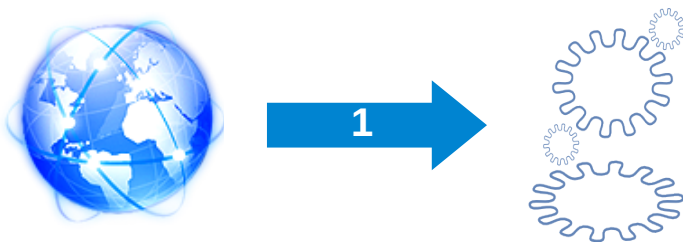
# Quick Gro

This document explains how to simply use Gromoteur for

- downloading a website
- cleaning the text
- exporting the text

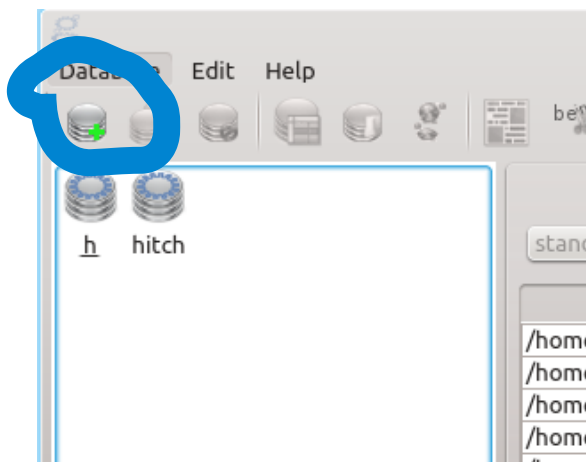
- Note that practically all fields in Gromoteur have a tooltip that explains the field's usage. Put the cursor over a field and wait !
- Regular expressions can come in handy for the full choices of configuration of Gromoteur. You can practice here, for example : <http://elizia.net/regex/> or at any website explaining Perl/Python style regular expressions.





## 1 Web to Gromoteur



Capturing web pages into a Gromoteur database in 10 steps:

1. Create a new base. Give it any name, then click OK. The base appears in the database panel.




2. Click on the button "g"  to open the web spider control window.
3. Click on the drop-down menu  and choose "New spider configuration". (Or, if you already have a configuration, choose it.)
4. Click on the button . The spider configuration wizard opens.
5. The  button allows to toggle between normal and expert mode:
  - normal mode : the essential choices

- expert mode: complete set of configuration options for the spider.
6. The configuration in 5 steps:
    1. Where do you want to put the collected pages?
    2. Where should it start crawling?
    3. Which way should it go?
    4. How many pages in?
    5. How do you want to name the configuration?



The details of these choices are explained in the next section of this tutorial.
  7. At the end, you click on “Finish” to return to the web spider control window.


8. Click on the start button  now to start the page collection process.

9. The page collection can stop for various reasons:
  - The maximal number of pages has been reached or the disk is full.
  - No more links are available to continue collecting.
  - The user has clicked on the  button to stop the spider manually. In this latter case, the user has to wait for each thread to finish its work, which is indicated by the button turning green again.


Now you can close the web spider control window by clicking on OK. You are back in Gromoteur’s main window now.

10. The main table shows now all the collected pages, one per row. Now you can

- restrict the view to specific pages using a filter 
- export the code source of the pages using the exporter 
- export the plain text into a single file or into separate files. See details in the “Export” section below.
- Systematically extract specific parts of the web pages into a separate column of the table

using the Field Selector . For example, in order to obtain the main content of a web page excluding advertisement or recurrent information. See details in the specific section below.

- Group pages in order to jointly export them or to analyze them statistically grouped together. {

- Do a statistical analysis using Nexico! 

## 2 Spider configuration for the collection of Web pages

### Basic rules of configuration

- proceeds by adding a few constraints, testing with few pages, and adding constraints little by little until the desired pages is obtained. Only then collect a greater amount of pages start collecting greater quantities of pages.
- When in doubt, leave default values.

The five steps of the configuration wizard have more choices when the expert button is triggered



### 1. Where do you want to put the collected pages?

4 possibilities :

1. *append new data*: add new pages at the bottom of the table even if the URLs already exist.
2. *overwrite an existing URL* : if the same pages are visited again, new data overwrites old data.
3. *erase all existing content* : erase the database.
4. *start from database* : use the stored collection of remaining URLs to continue the crawl (only available if the base already contains links to be explored).

When in doubt, choose 3 (*erase all existing content*)

### 2. Where should it start crawling?

3 possibilities :

1. Start with one (or more) web addresses (several URLs can simply be entered, separated by spaces): Put the URL(s) in the line.
2. Provide a file containing the URLs to visit, one URL per line.
3. Start crawling with a search engine query. Gromoteur uses Bing.
  - If you have many queries to do (say, more than 10 per day), please created your own account at [www.bing.com/developers/createapp.aspx](http://www.bing.com/developers/createapp.aspx) in order to not get in the way of other users of Gromoteur.
  - Put the keywords in the corresponding line. To see all options visit: <http://onlinehelp.microsoft.com/en-us/bing/ff808421.aspx>
  - Test whether the query produced the desired results by pushing the « try Bing » and then the « Firefox » button.
  - Constrain the outcome by giving the desired URL by means of a regular expression in the “follow only” line.

### 3. Which way should it go?

In *normal mode*, you can only give conditions on the pages that you want Gromoteur to stay on.

Only download page if:

URL matches

URL doesn't match

page contains

page is in

level from  to

use links only if page is OK

If the box is checked, Gromoteur will compare each page with the given conditions before adding the page to the database. It is possible to constrain the URL (in the image above, we want only pages from « popsci.com » whose URL contains letters (\w+) before the word “article” and a date starting with 201. In this way, we can provide positive or negative

constraints on the URL, the plain text content, or the language of the page (test based on trigrams, see <http://www.elizia.net/languageDetector/> for further information). Moreover, it is possible to constrain the level, starting from the initial URL, on which Gromoteur should collect the data: Level 0 corresponds to URLs that were given or obtained in step 2 (“Where it should start crawling”), level 1 are the links that were found on these pages, etc. The level distinction is not completely tree-like of course, as it is possible to have back-links to previously visited pages with higher levels. You can also decide whether Gromoteur should “use links only if page is OK”.

- If this choice is checked: The links will be explored further only if the page fulfills the conditions.
- If this choice is unchecked, Gromoteur will use the extracted links on the page even if the page itself failed the conditions and was not added to the database.

In *expert mode* a few more choices are available:

- The path: either visit the collected links in the order of the level of the pages: first level 0, then level 1 (breadth first), or dive deep by always following the first (given or found) link (depth first).
- It is possible to provide specific conditions on when the link should be added to the list of links to visit.
- It is possible to include pdf files when encountered during the crawl.
- It is possible to specify the page encoding if Gromoteur doesn't make correct guesses.

#### 4. How many pages in?

Here we give the number of required pages. It is useful to start with a small number (10 to 30 pages). It is also possible to indicate this restriction in number of sentences (estimated for simple segmentation based on punctuation). Alternatively, it is possible to restrict the collection by the size of the database or by the remaining space on the hard drive used for storing the data.

In *expert mode*, it is also possible to restrict the number of sub-domains to visit (xxxx.domain.com), which can be useful to obtain complete blogs on a blog website.

This configuration page allows for other parameters, too:


- Avoiding spider traps by simple heuristics: Limiting the speed with which Gromoteur visits each server and by trying to change the order between the links to be visited so that each server is visited infrequently.
- To follow or not to follow URL redirections (i.e. links that automatically change to a different URL when visiting)
- Setting the time-out: After how long should it try again or give up waiting for an answer from the server?

#### 5. Naming the configuration.

In *normal mode*, you only give a name to the configuration that you just created.

In *expert mode*, you can also decide:

- How many threads should be started in parallel by the system. The best number depends on the number of cores of the computer, the connection speed and the speed of the web server. Just try some values.
- Proxy settings (*http*, *https*, *socks5*).
- How Gromoteur should present itself, as Gromoteur or under a different identity?
- Should we be polite and obey to the server's *robots.txt* file? See [http://en.wikipedia.org/wiki/Robots\\_exclusion\\_standard](http://en.wikipedia.org/wiki/Robots_exclusion_standard) for further information.

Close the configuration by clicking on “Finish” and start collecting with the  button.


### 3 Cleaning Web pages



Most Web pages contain parts that are not interesting for linguistic analysis: The generic parts of each page (the title of the newspaper, links to other sections, legal information, ...) and advertisement. Often, the desired parts of the Web pages have specific HTML markups. Gromoteur can discover these parts and extract the interesting segments.

1. In the main window, open the base to be used, select a page that contains the desired parts that you want to extract:

27	<a href="http://liberation.fr/monde/2014/02/19/en-direct-l-ukraine-ouvre-une-enquete-pour-tentative-de-prise-illegale-du-...">http://liberation.fr/monde/2014/02/19/en-direct-l-ukraine-ouvre-une-enquete-pour-tentative-de-prise-illegale-du-...</a>	19/0...	1
28	<a href="http://liberation.fr/societe/2014/02/19/le-long-et-difficile-combat-pour-un-statut-du-stagiaire_981411">http://liberation.fr/societe/2014/02/19/le-long-et-difficile-combat-pour-un-statut-du-stagiaire_981411</a>	19/0...	1
29	<a href="http://liberation.fr/societe/2014/02/19/des-salaries-de-l-usine-depalor-retiennent-trois-dirigeants_981584">http://liberation.fr/societe/2014/02/19/des-salaries-de-l-usine-depalor-retiennent-trois-dirigeants_981584</a>	19/0...	1
30	<a href="http://liberation.fr/economie/2014/02/19/les-syndicats-de-fagorbrandt-decus-apres-une-reunion-a-bercy_981514">http://liberation.fr/economie/2014/02/19/les-syndicats-de-fagorbrandt-decus-apres-une-reunion-a-bercy_981514</a>	19/0...	1

2. Click on  to open the Field Selector.
3. Select the parts that you want by clicking on them. They will turn blue:

manifestent pas.

Une centaine de salariés de l'usine de panneaux de particules en bois Depalor à Phalsbourg (Moselle), qui ont perdu leur emploi après un incendie accidentel en juillet, faisaient le siège mercredi soir de leur usine et retenaient trois membres de la direction, a-t-on appris de source syndicale.

Le directeur général, le directeur technique et la directrice administrative et financière de l'usine sont retenus dans les locaux administratifs depuis 9h30 et «invités à rester» sur le site aussi longtemps que le PDG de l'entreprise, appartenant au groupe helvétique Krono, «ne sera pas là pour débloquer la situation», a expliqué le représentant de la CFDT Michel Beltran.

En décembre dernier, la direction de Depalor avait annoncé aux salariés la fermeture du site après l'incendie accidentel et justifié sa décision par la «situation

tag : DIV     id : article-body     class : article-body mod

Une centaine de salariés de l'usine de panneaux de particules en bois Depalor à Phalsbourg (Moselle), qui ont perdu leur faisaient le siège mercredi soir de leur usine et retenaient trois membres de la direction. a-t-on aoris de source syndica

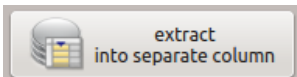
4. Below the web page, you see the criteria used to mark this page:

tag : DIV     id : article-body     class : article-body mod

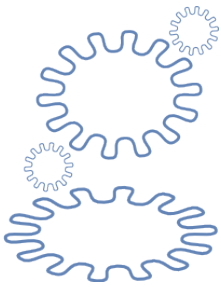
5. Try out whether these criteria really work for what you want to extract by clicking around your database using the “previous” and “next” buttons.
6. Once the selection is OK, choose a textualisation (or just leave the default *standard*) and provide a name for the column to receive the data from the desired parts:

Textualization Name:

Column Name:

7. Click on the  button to start the extraction of all the pages of the current database.
8. If you do not have another part to extract, you can now close the Field Selector and return to Gromoteur’s main window. You can see the new column(s) in the very right of the table.




## 4 Export

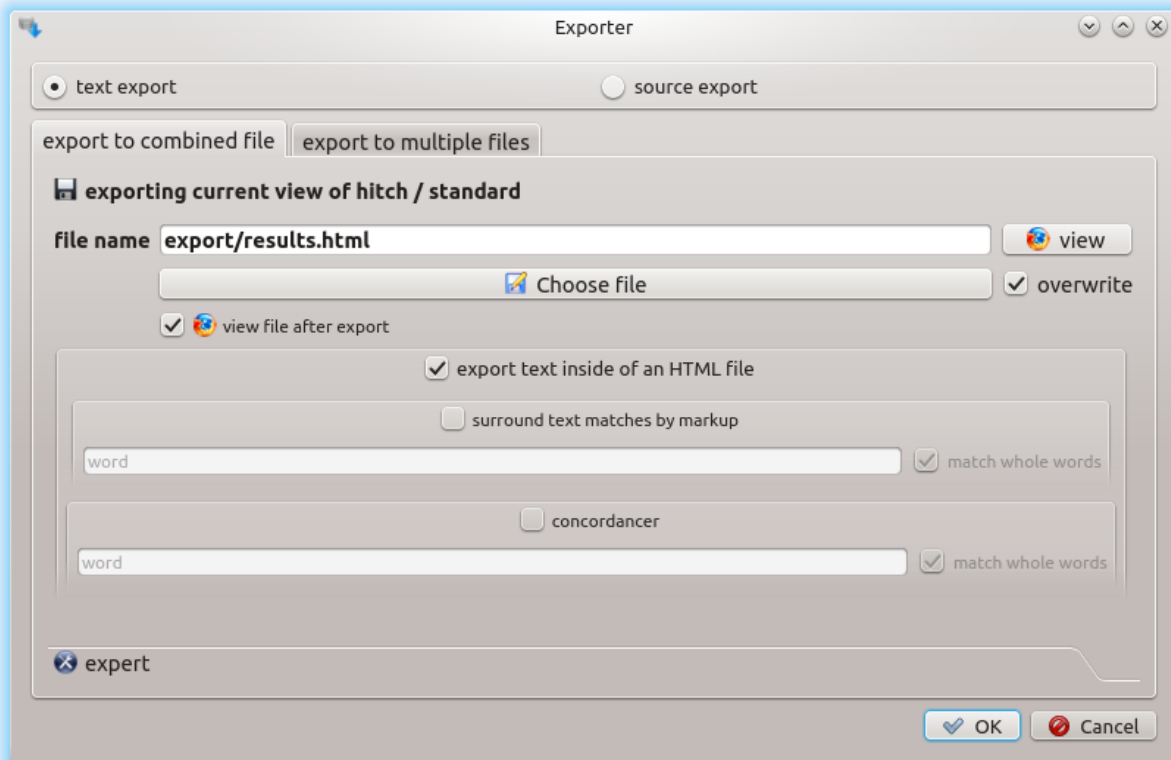



Gromoteur allows to export a database into text files.

Basically, there is the choice between exporting HTML source code or plain text.

In order to export plain text, proceed as follows:

1. Select the content to be exported in the main window:
  - a) Choice of columns: Right click on the header of the main table allows to hide or to show content. The  menu allows to choose the desired columns.
  - b) Choice of lines: With the “current filter”  tab, you can select the desired lines.
2. The button  (or the menu *Database*, then *Export*, or also the shortcut CTRL-E) opens the exporter that allows to export the database content in different types of files:



3. The default choice is that the text content is made into a single HTML file, which opens when the OK button is clicked. The exporter uses a template that can be chosen in the “expert”  options below.

- We can get a color markup on a word by putting it instead of “word” in the line below “surround text matches by markup” after checking this option.
- It is also possible to export the concordances of a word by putting the word in the line below “concordancer” after checking this option. The result looks like this:

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 Score 72205 359 9 Dear Congress: Why Are You So Anti-Science? 73263 359 10
<a href="http://www.popsci.com/science/article/2013-05/how-avoid-meeting-neanderthals-fate">http://www.popsci.com/science/article/2013-05/how-avoid-meeting-neanderthals-fate</a>		
21.	What Modern Humans	Learn From The Neanderthals' Extinction   Popular Science
22.	What Modern Humans	Learn From The Neanderthals' Extinction   Popular Science Login/Register Newsletter Subscribe
23.	science 73784 What Modern Humans	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	we avoid meeting the Neanderthals' fate? That depends on what you think wiped out these early
25.	Extinction 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	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	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	'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 GodLiesComedy 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 S
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 Barbecue	Can Do Division Compute Longshots And Take Square Roots Meet Viewed Science The Week In Numbers

- For the two preceding choices, regular expressions can be used.
- The second tab “export to multiple files” allows to obtain one file per line of the table. The file names can be the URL’s or a simple number.



To export source code proceed as follows:

- Choose the lines that you want to export, the selection of visible columns has no influence on the source code exportation.
- Start the exporter.
- Choose “source export” above.
- Choose between unique or separate files by opening the corresponding tab.
- Click OK.

## 5 Files to Gromoteur



Two ways of importing files:

- A whole folder 
- A tab-separated (csv-type) file 

The folder can contain text files (utf-8) or pdf-files. All will be converted to simple text and imported.

The tab-separated file puts the last column as text column, and the preceding columns, space-separated, as title column.