

URL Extractor 3.3

User Guide



Tension Software - We Make Software for Mac - Pomola.com

URL Extractor © 2006-2011 - Tension Software all rights reserved

Every effort has been made to ensure that the information in this manual is accurate.

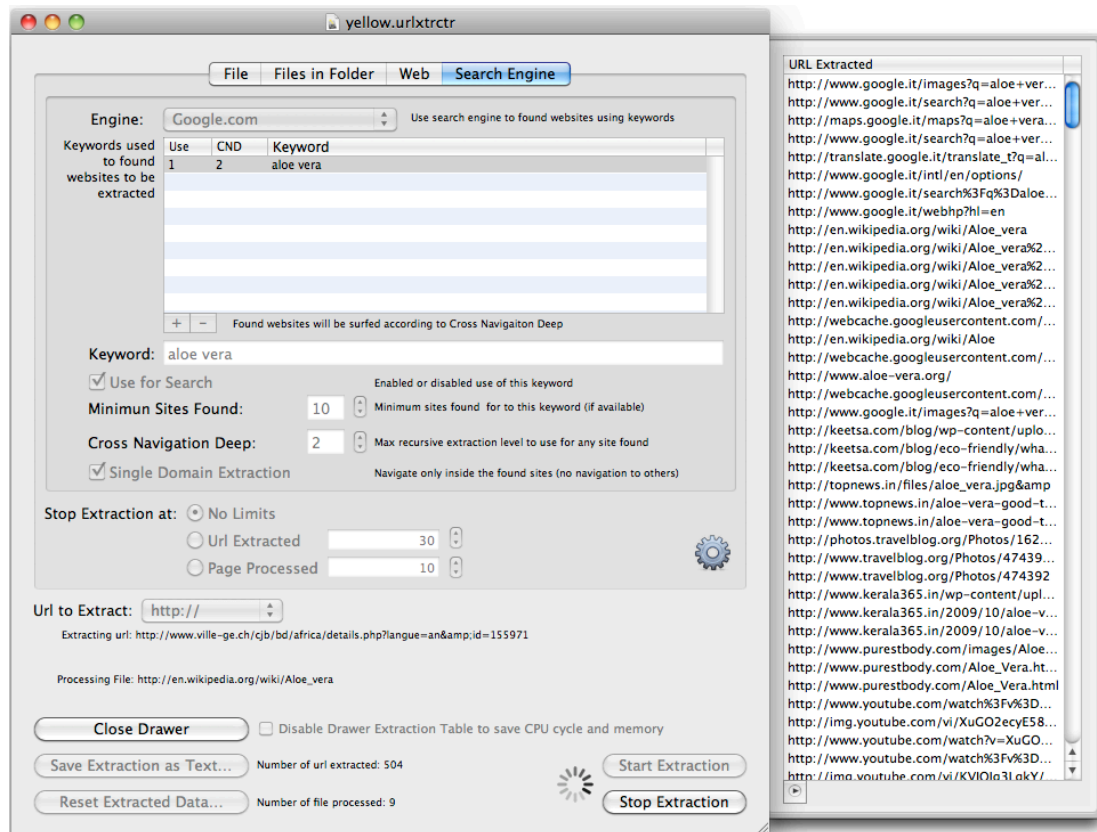
Tension Software is not responsible for printing or clerical errors.

Other company and product names mentioned herein are trademarks of their respective companies.

Welcome to URL Extractor

Extract URLs from Everywhere

The Cocoa application to extract emails address, web address and URLs in general from files on Hard Disk, from the Web and also starting navigation from search engines keywords.



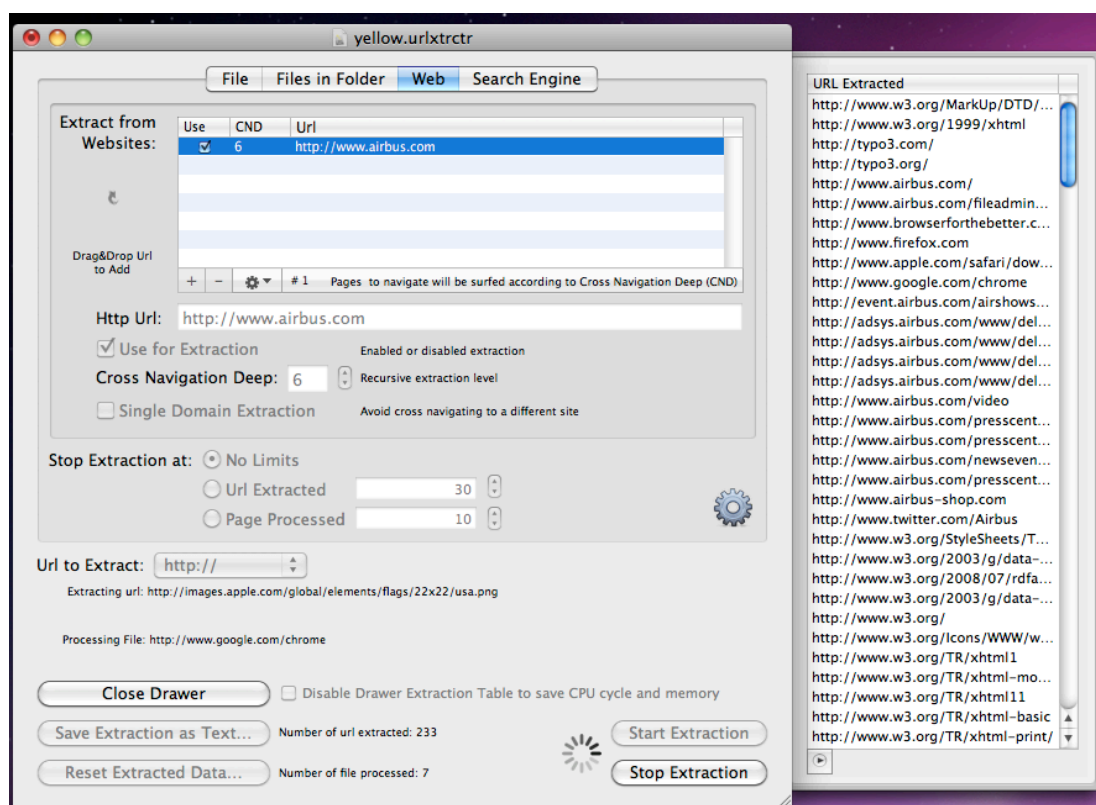
Getting Started with URL Extractor

URL Extractor is Really Simple to Use

The interface is simple and intuitive and you can save the operation settings, in URL Extractor documents on disk.

The advantage is you can iterate over time the same extraction (as example the same keyword search, once a week) without having to reinsert all the parameters, just reopening the same document from disk.

Extracting from files on your hard disk, you can process a single files or thousands and thousands of them, in a single shot, in just few seconds with an amazing multithreading technology, all Cocoa native.

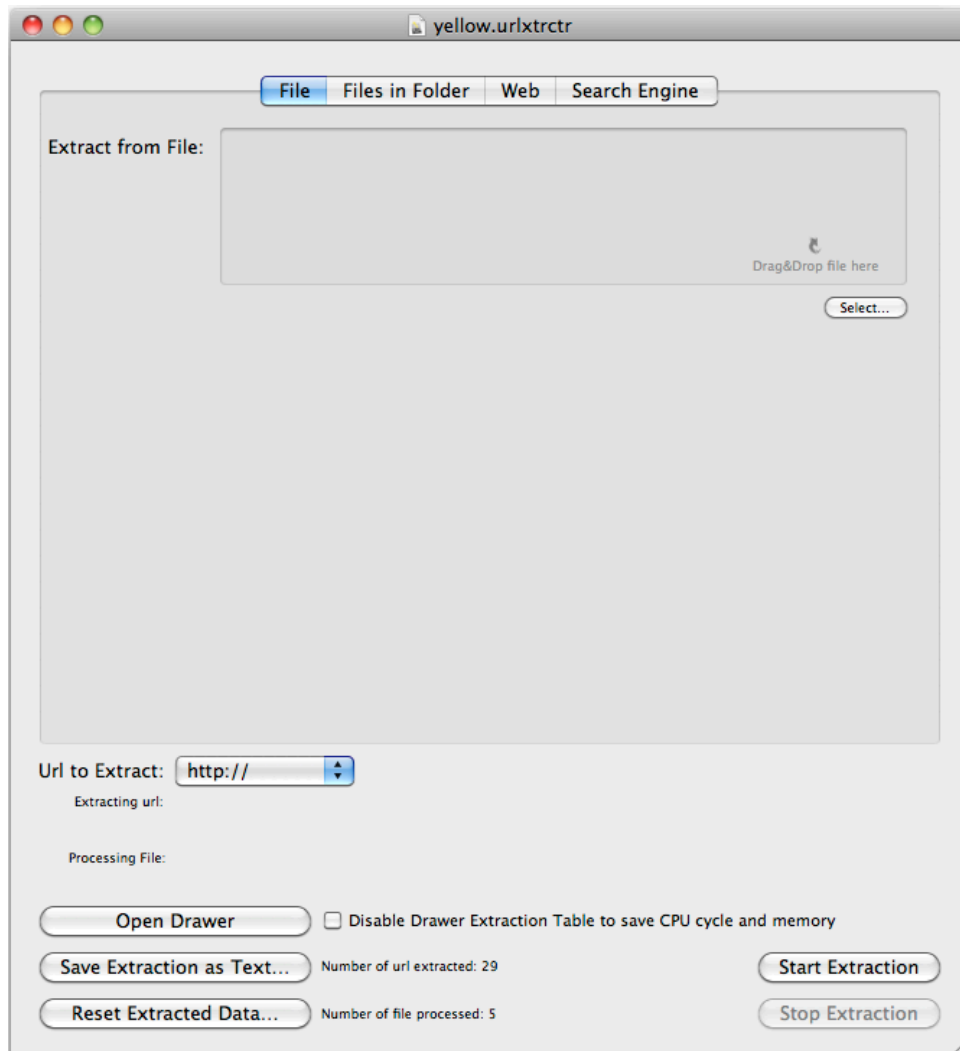


Extracting from 'Web' you specify some starting web pages and URL Extractor extracts the urls from the indicated pages, then it navigates in the linked pages (cross navigation) extracting the urls and so on in a process that can extract a gigantic amount of data, starting from a single web page.

Using 'Search Engine' you specify some keywords, URL Extractor will use Google to make search using these keywords and analyzing the resulting pages and the sub sequential linked pages to search for the url types you want to extract

Extracting from a File

Open a new document or use the one opened for you at startup.
Using the tab in the upper part of the window select "File"
Using the button "Select..." select a file from disk.
You can also drag and drop a file from the finder in the Url Extractor window
Select which kind of url you want to extract , "http://" or "ftp://" or a mail address
prefixed with "mailto:" or others url or simply email embedded inside any content
Press the "Start Extraction" button.
Done.



All the url are now ready to be saved on disk for you in text format.
This was to process a single file.
What about if you have to process 10.000 or more files nested inside different
folders?
It's easy almost as processing a single file.

Extracting from all the Files in a Folder

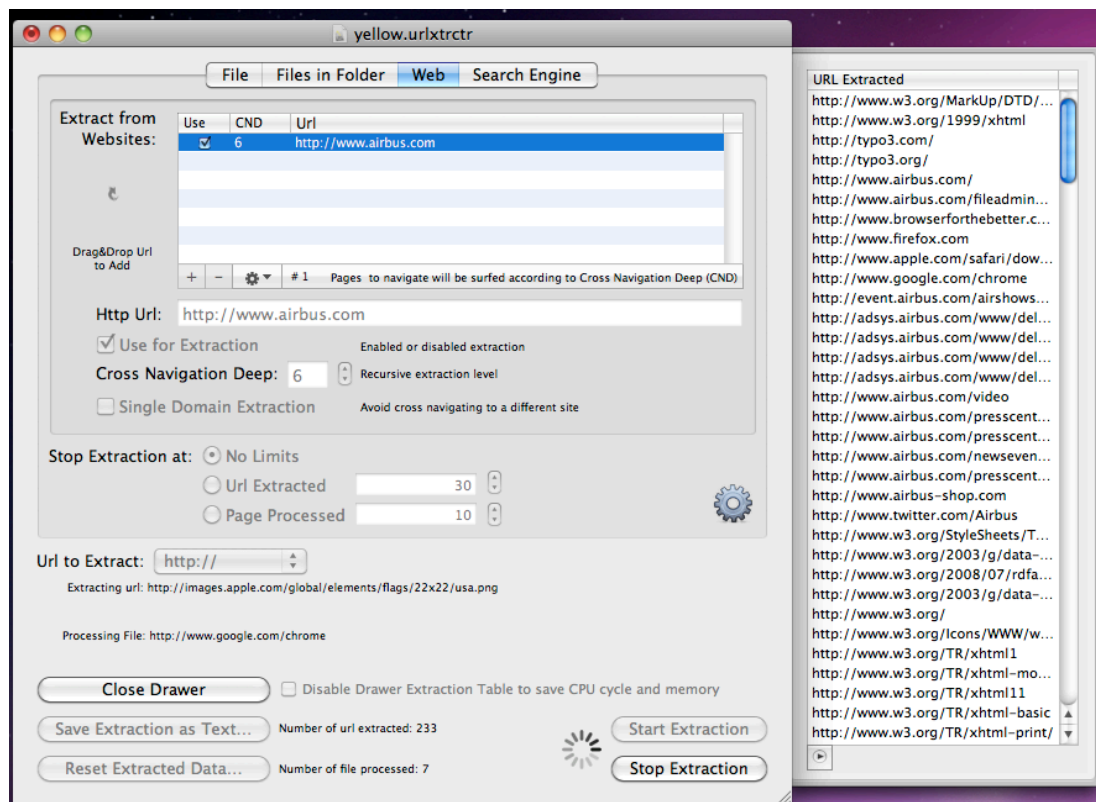
Select from the window the "Files in Folder" tab.
Use the "Select..." button and select the root folder containing all the hierarchy of
folders and files you want to process.

You can also drag and drop a folder from the finder in the Url Extractor window
You can further customize the file to process selecting to extract from "any" or selecting the kind of file to process specifying the extension of them.
Specify as in the previous example the type of url to extract.
Pressing the "Start Extraction" button
Done.

And now the more powerful part of Url extractor, extracting from the web:

Extracting from the Web

Select from the window the "Web" tab
Use the '+' button and add a starting web pages
You can drag and drop a URL of a web page from Safari



Indicate the Extraction level you want the process go deep using 'Cross Navigation'
Be sure the 'Use' check mark is checked to have the Url used for extraction starting
Indicate when to stop the process, 'No Limits' never stops, or 'Url Extracted' stops when the extracted URLs reach the indicated number, 'File Processed' stops when the processed web pages reach the indicated number.
Select the type of url you want to extract
Press 'Start Extraction'
You can stop manually the extraction using the 'Stop Extraction' button

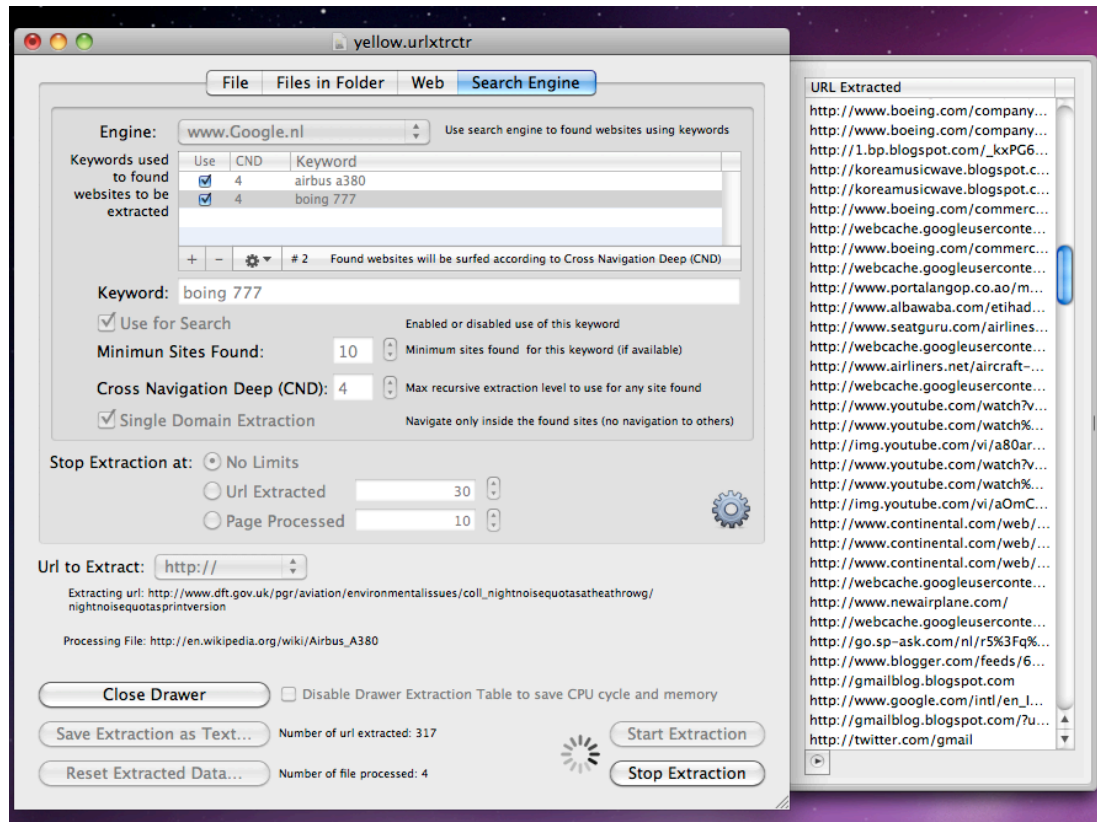
Another way to extract from the web but starting from keywords is extracting from 'Search Engine'

Extracting from Search Engine

Select from the window the "Search Engine" tab

Use the '+' button and add a keyword

Indicate the Cross Navigation Deep you want the process go deep using 'Cross Navigation' Be sure the 'Use for Search' check mark is checked to have the keyword used for extraction starting



Minimum site found indicates how many web site at minimum to search for a specified keyword (Url Extractor will query the search engine to have at least the indicated value of sites found)

Single Domain Extraction navigate only inside the found sites (at Cross navigation deep level)

Indicate when to stop the process, 'No Limits' never stops, or 'Url Extracted' stops when the extracted URLs reach the indicated number, 'File Processed' stops when the processed total web pages reach the indicated number.

Select the type of url you want to extract

Press 'Start Extraction'

You can stop manually the extraction using the 'Stop Extraction' button

Common Things for all Extractions

You can watch the progress of the operation and also the total of files and url processed.

Opening the 'Show Drawer' you can see the url as they are collected 'live'

The operation is performed in multithreading and you can stop it using the "Stop Extraction" button at any time.

All the url extracted are ready to be saved on disk for you in text format (fully enabled only in licensed version).

You can save the extracted url on disk as text file, every url separated by a new line inside a text document.

You can save also the document (not the url extracted), containing all the settings used for the extraction, as a URL Extractor document.

Reopening it you will have all the setting ready for a new extraction, useful if you often extract with the same setting and from the same folder where you save files over time.

Using URL Extractor let you do in seconds what takes days or months to do manually.

URL Extractor employes the latest Apple technology in OS X and provides a software you will use now and in future because is entirely based on the best solution on the planet without compromises, Objective-C and Cocoa, even fore core extraction algorithms implemented using Cocoa multithreading.

URL Extractor Reference

Application Organization

URL Extractor is a standard Cocoa Document based application.

A window can be saved on disk as a URL Extractor document with all the settings used for the extraction and reopened at later time.

All the settings saved for the extraction will be available again without the need to reinsert it.

If you need to use often a extraction setting, consider saving a document with these settings inside. You will avoid to have to reinsert data anytime as you have to do in a typical utility application without document saving capabilities.

Files, Files in Folder, Web and Search Engine

Extraction operation can be performed on:

A single file on your Hard Disk

All the files contained in a folder hierarchy on your Hard Disk.

All the web pages indicated in a list and all the followed linked pages to a specified deep (Cross Navigation Deep)

Search Engines using selected keywords, the keywords generates a list of web links used for starting a cross navigation search

Selecting inside the document window the tab "Files" or "Files in Folder", "Web" or "Search Engine" you select which type of extraction to use.

In case a single file is selected, it will be processed. It must be readable as file text.

In case the "Files in Folder" is selected, all the files inside the folder, also files inside nested folder at any level, will be processed to perform the extraction. They must be readable as text files to be used.

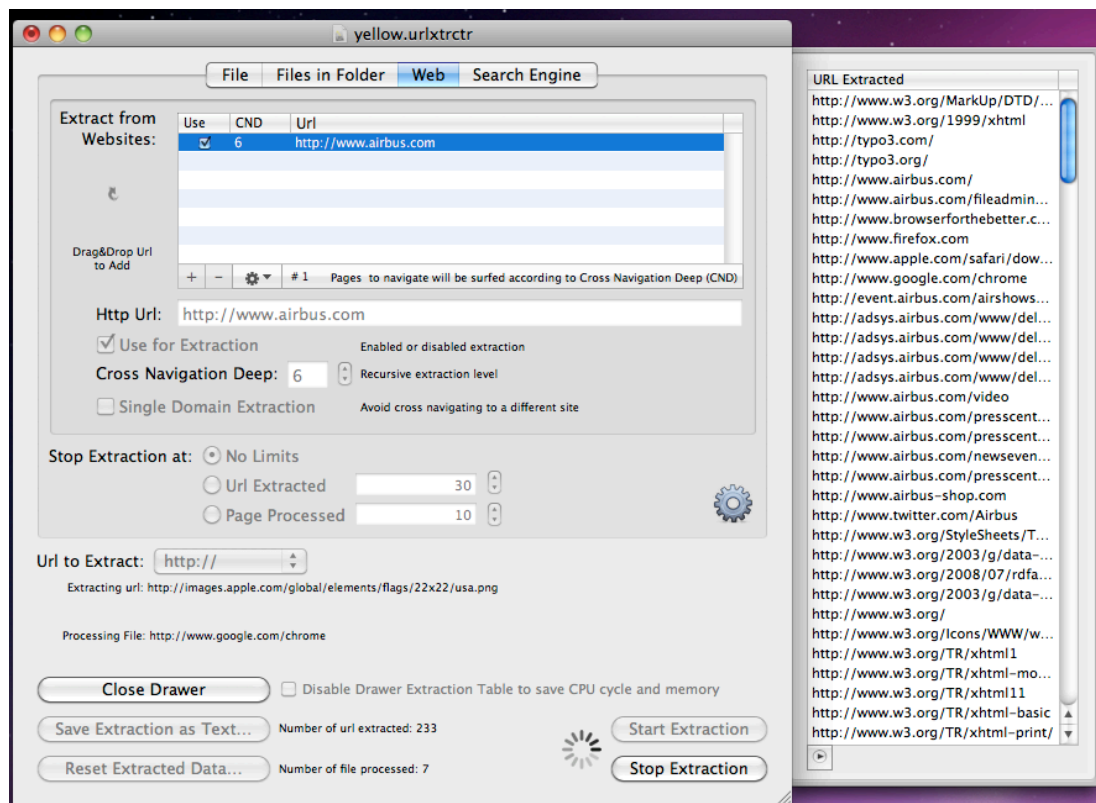
In case "Web" is selected you can specify a list of web address. This list will be used as starting point for cross navigation. Url Extractor will navigate these pages and all the (inside these pages) linked pages in background using multithreading. This can produce amazing result, starting form a single web page address Url Extractor can navigate hundred and hundred of web pages extracting a large amount of url, all with a single pressing of a button

In case "Seach Engines" is selected you can specify a list of keywords. This list will be used as starting search . For any keyword the extraction will generate a list of web page addresses, at minimum the value specified in 'Minimum Site Found' and Url Extractor will navigate all these pages and all the (inside these pages) linked pages.

All in background using multithreading with just a single click. This is powerful as much as the Web' extraction or more. Using a keyword Url extractor will find all the related urls on the web working for hours unattended

Show Drawer

Show Drawer let the user see url 'live' as they are collected in a table. To fully enable urls collection, Url Extractor must be licensed.



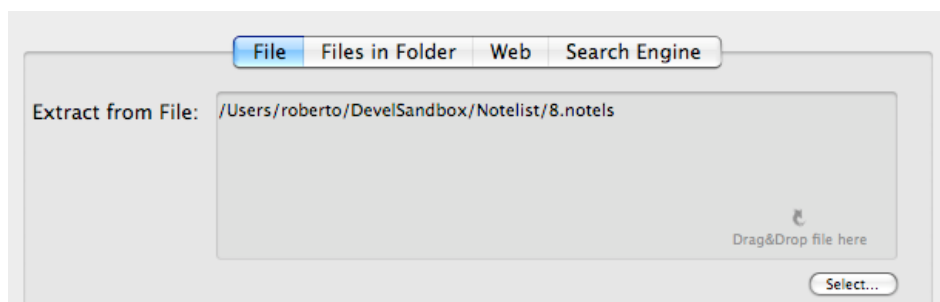
To launch a url double click it in the drawer or select it and use the launch button (at the base of the drawer table).



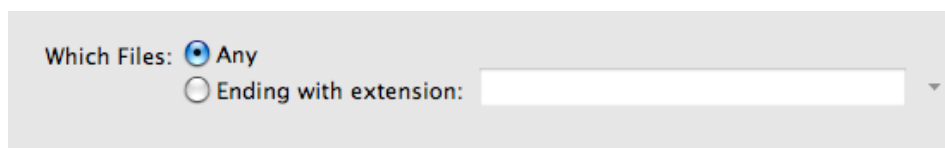
It will be opened using your default application for the url. (Example: Safari for a 'http://' page in case you maintained the default settings)

Which Files to Extract

If you select a specific single file to extract from, no further specification are necessary.



In case you select "Files in Folder", you can restrict the criteria specifying which file to analyze.



Which Files: Any
 Ending with extension:

You can specify many extension inside the "Ending with Extension" field and only the files matching this extension will be used to extract url.

This can speed up operations a lot if you are using a folder with a lot of files, many of them useless to extract url, such as graphical files inside web sites and so on, if these files have an extension different from the specified inside the "Ending with Extension" field, they will be skipped saving time during the extraction process.

Using Web extraction, Url Extractor try to decide by itself the correct type of file to navigate, no input required.

Selecting a File of a Folder

To select the file or the folder to work on you can press the 'Select...' button in the dialog or you can drag a file or a folder from the finder directly in the area inside the URL Extractor window.

The file or the folder will be selected as the one to process.

Inserting a Web Page to Navigate

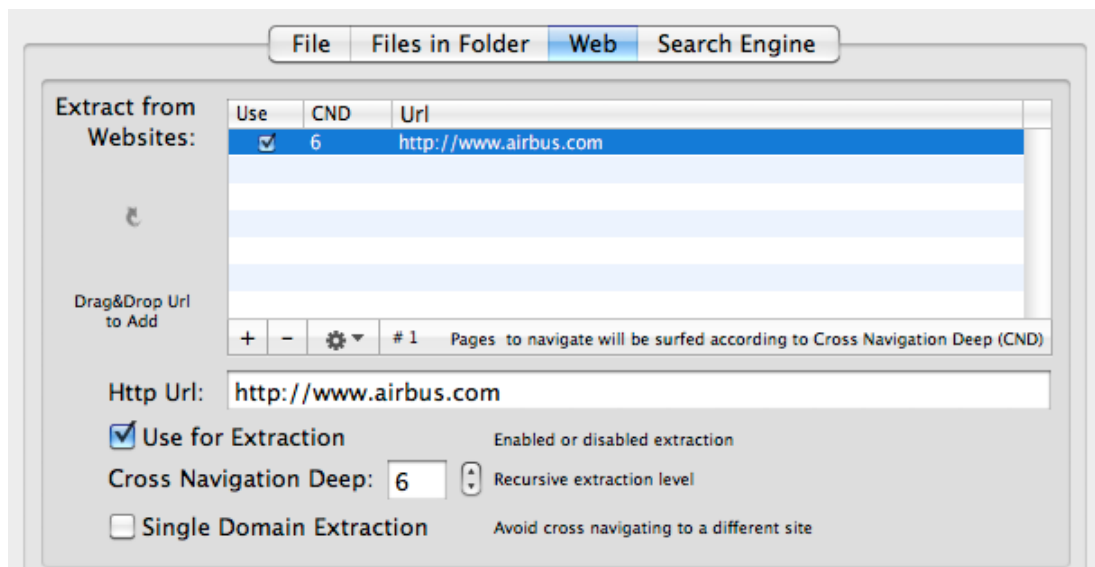
A web page to navigate can be inserted editing the url or drag and dropped from Safari or other programs

Web Extraction

Url Extractor can extract from the web starting from a list of pages and for each one cross navigating all the linked pages and extracting the indicated url type for any visited page.

Considering almost any page on the web is linked by others, this technique can let you extract thousands and thousands of url (email or others) from the web starting form a single web page!

To indicate the web page to use for starting the extraction process, you can edit them or just drag and drop them from other program such as Safari or Firefox.



Be sure the web pages start with the right protocol tag 'http://' as example '<http://www.apple.com>' when manual editing it.

You can indicate as many pages as you like in the list.

Cross Navigation Deep in web extraction is used to perform cross navigation extraction, as example if you indicate a value of 3, Url Extractor starting form the indicated page will navigate ALL the linked web pages for 3 level of link in deep, this will cause that starting from a single page, you can navigate and extract a lot of pages if there are many linked pages in any of them.

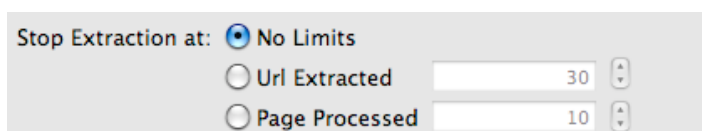
You can have web page that can be in the list and not used for extraction. Use the 'Use' check box to toggle between use and not use.



If 'Use' is unselected, the page won't be used for extraction and it won't be used also for cross navigation (extracting links and using linked pages)

Url Extractor allows you to specify when to stop web extraction. Considering that cross link navigation allows the software to extract (if starting with a page with many links) a great amount of data. This was necessary to limit the amount of data that the program can download from the web when not requested and the time of the total extraction process.

You can specify the following stop extraction events:



No Limits: The program will stop extract when it run's out of page to extract or because the user pressed the 'Stop Extraction' button.

Url Extracted: The program will stop to extract when the extracted URLs number reach the specified value.

File Processed: The program will stop to extract when the processed file number (web pages) reach the specified value.

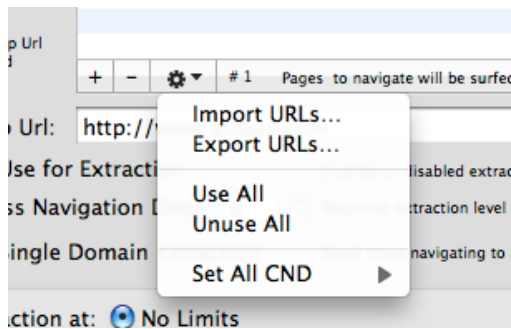
Selecting 'Show Drawer' shows the table where the urls are collected as they flow inside the data.

After extraction, extracted urls can be saved on disk in text file format to be used for your purposes.

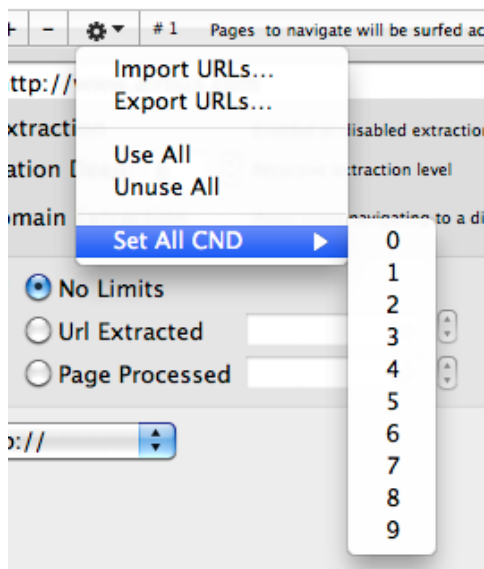
Importing URLs

The user can import a list of URLs from a file on disk

Select 'Import URLs' from the file menu or select 'Import URLs' from the popup command.



Select, in case the file to import has more then one column, which column to import in the successive panel.



Exporting URLs

Value in the URL table can be also exported to a file on disk

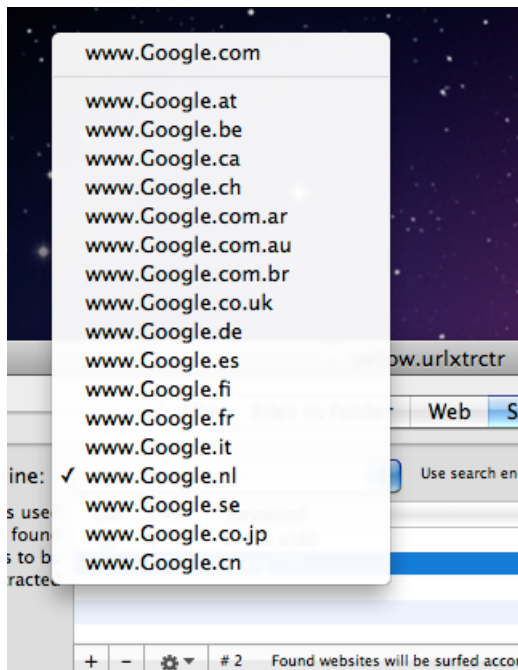
Simply select 'Export URLs' from the file menu and name a new file to save the full list on disk.

A single field (one column) file will be created with as many rows as the URLs to save

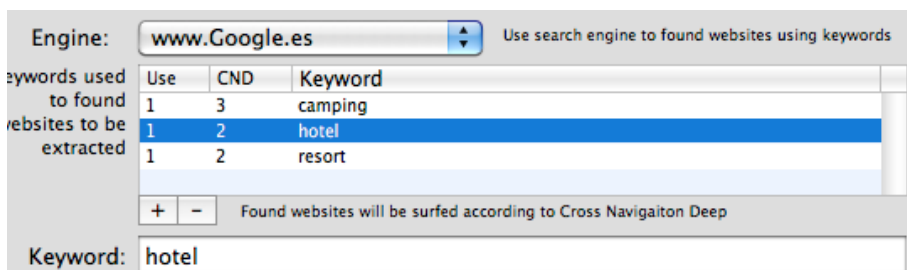
Search Engine Extraction

Search engine extraction is very similar to web extraction. It just adds a previous step, it search on a search engine the pages associated with a search using a specified keyword, then it starts web extraction starting from all the web pages found.

The user can specify which search engine used for extraction from the popup with the list of all the usable search engines



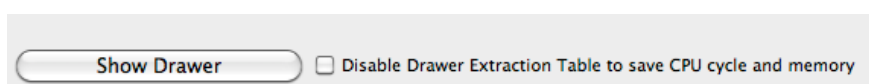
The user can specify a list of keywords, search for any keyword the associated web pages and start a cross navigation extraction from any web page (keywords can be imported from an external file too) .



The software let analyze and extract a gigantic amount of data just starting from a single keyword. Selecting the corrects keyword the user is able to search the net navigating the sites associated with the keys the user consider interesting. The software let specify the minimum web site to found (and navigate) for any keyword.



Url Extractor queries the search engine until the requested number of sites are found (associated to a given keyword), then it starts cross navigation and Urls extraction. Selecting 'Show Drawer' shows the table where the urls are collected as they flow inside the data.



After extraction, extracted urls can be saved on disk in text file format.

Url to Extract

The Url to Extract popup menu is used to specify the kind of url to extract.

You can specify:

"http://"

"ftp://"

"mailto:"

"feed://"

"telnet://"

"telnet://"

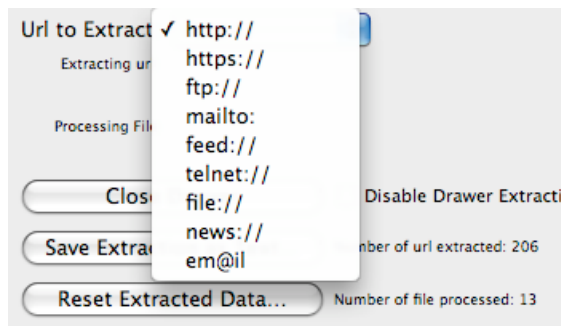
"file://"

"news://"

and all the URL starting with this prefix will be extracted or also

m@il

and all the generic emails embedded inside any content and also without any prefix will be extracted. They are recognized just by the '@' presence.



Start the Extraction

Pressing the "Start Extraction" button starts an extraction process.

It is performed in a separate thread to let the application continue to react perfectly at the user input.

Searching and extracting on the hard disk thousands of files in a old fashion mono threading task, will cause the program to "freeze" for all the time of the process, using the multithread feature, native in cocoa and in unix, makes URL Extractor, a joy to use!

The best part is that the multithreading feature is native and implemented in a "bullet proof" way in Cocoa and URL Extractor uses it in a simple and elegant way.

Perform an extraction on a folder with lot of files inside to see it in action.

Another way to see it is using web extraction. Web navigation and extraction is performed in a separate thread and Url Extractor reacts perfectly to your input even during a heavy load downloading and processing hundred of web pages.

Stop Extraction

Pressing the "Stop Extraction" button stops the extraction at any moment sending to the thread performing the extraction a message to stop. All is performed in a clean way, nothing is interrupted in a brutal way.

In case you restart the extractions, it is restarted from the begin again, so to extract all the files inside a folder (nested subfolder included) you need to perform an extraction completely from the begin to the end, when the extraction process searched in all the files, it stops itself.

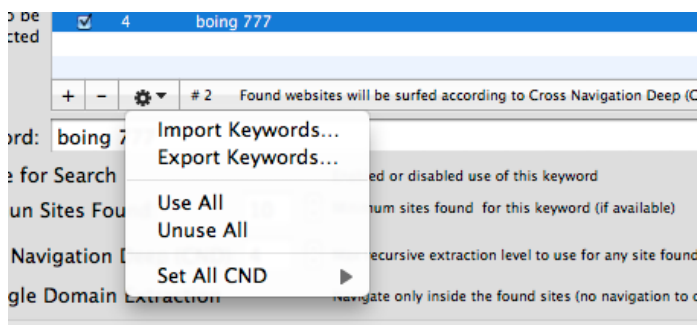
In case the extraction process takes too much (probably because you are extracting form a folder with inside a lot of files) you can use the "Which Files" option to specify the extension of the files to analyze and speed up the process.

Using Web extraction you can specify additional event that interrupts the extraction

Importing Keywords

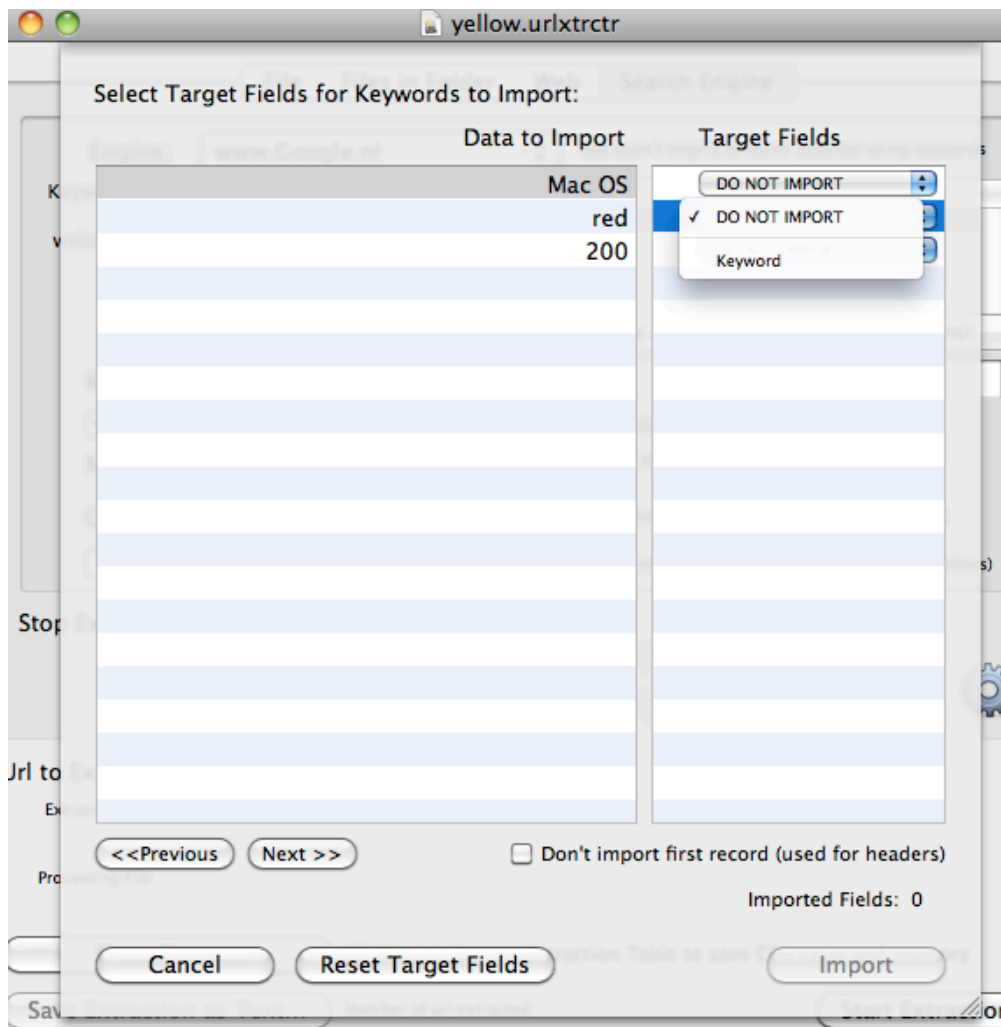
In a way similar to the URLs also keywords can be imported from file on disk to the application

Select 'Import Keywords' from the file menu or select 'Import Keywords' from the popup command.



Select, in case the file to import has more then one field, which field to import from in the successive panel.

In case the text file has just one (column) field, then select the single field present in the file. All the records will be imported (for the selected field)

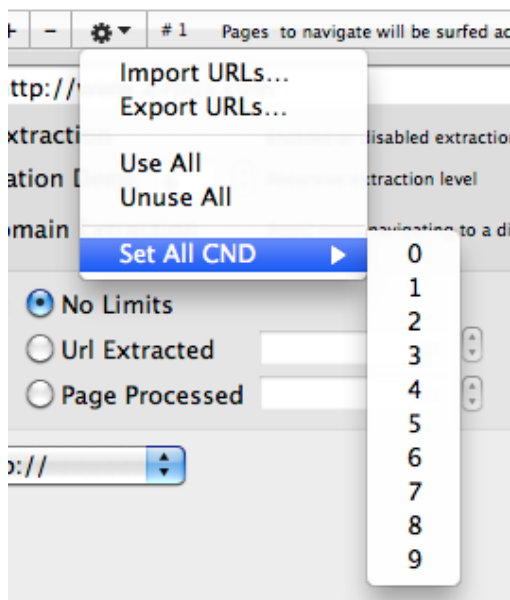


The panel allows you to browse the file you are importing to see its content and select the right field to import in the application to 'Keyword' and leave a 'DO NOT IMPORT' for the other fields (if there are other fields in the text file you are importing from).

All the value of the selected field will be imported from the file inside the application Press the 'Import' button to complete the import

The 'Import' button becomes enabled when you have selected the field to import from

Using the popup command at the bottom of the table you can assign a 'Cross Navigation Deep' (CND) to all the imported values in a single command You can also set to select all for use or as not-use



Exporting keywords

Value in the Keywords table can be also exported to a file on disk

Simply select 'Export Keywords' from the file menu and name a new file to save the full list on disk.

A single field (one column) file will be created with as many rows as the keywords to save

Speed

Some example using a old low level Mac, a Mac Mini with a 1.42 GHz Power PC and 512 MB of Ram:

It took 3 seconds to extract 655 email address form a folder with subfolder and 121 .html files (the Ending with Extensions: "html htm" was used)

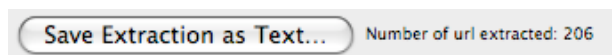
Removing the Ending with Extension and performing a search an any files gave this result: 4 seconds to extract 655 email from 218 files. If there are a different percentage of non significative files (containing url in text format) inside your folder the result may be a lot different. From a general point of view we can say that if there is a significative percentage of non text file it is better to use the "Ending with extension" option.

Using a latest generation Intel iMac or Macbook, URL Extractor can literally fly.

Web extractions and Search Engine extraction depends a lot of the internet download speed you have at your location and how fast web pages are loaded. It may change a lot depending of your local connection and all the sites you analyze.

Save Extraction as Text

After an extraction all the url extracted are available in memory to be saved on disk. Pressing the "Save extraction as Text..." button you can choose where to save. the url will be saved as a text file, with every url separate by the other by a newline. You can import this file easily in any software such as database spreadsheet, text processor and so on. A double click on it generally opens it in TextEdit (it depends of your computer setting)



Please note that you must purchase a license to save the full extraction on disk. Unlicensed copy are limited just to evaluate the program.

Reset Extracted Data

The "Reset Extracted Data" button resets all the extracted data. After that no url are available in memory to save on disk as text. To have them again you have to extract it again using the "Start Extraction" button.

Live Extraction

During extraction the extraction you a visual feedback of the process. The following data are showed as they are processed:

the url extracted

the file analyzed

the total number of url extracted

the total number of files (web pages for web extraction) processed

General Preferences

General

Specify the action to do at startup

New Document create a new document at startup

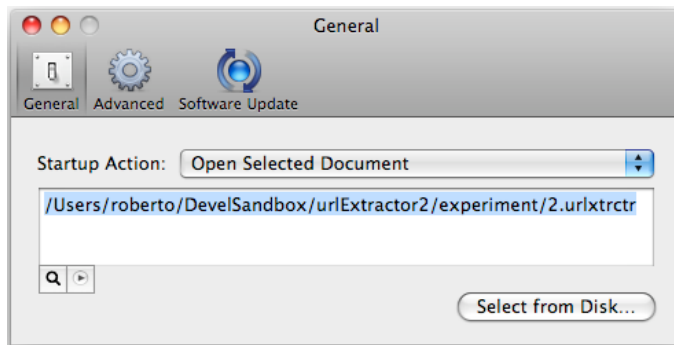
File browser open the dialog to select a URL Extractor document from disk to launch

None performs nothing at startup

Open Selected Document allows to select a document and have it launched automatically at startup, this is a standard way to operate in case you use just a single document as the main repository of your notes or use often the same document.

To select a document you have 3 way:

- Pressing the select from disk button and select it from the open dialog that will follow
- Dragging it from the finder or dragging it using the proxy icon (the icon in an opened document in the title bar)
- Editing it in the edit field by keyword (the hard way)



At any successive relaunch URL Extractor will execute the option selected. The additional two buttons at the foot of the edit field let you:

- Show in the finder the selected file
- Test open the selected file as it will be done at the next application launch

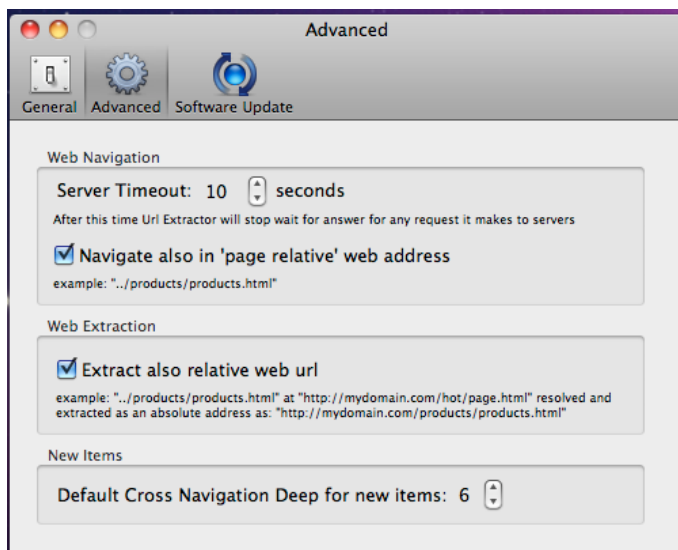
Advanced Preferences

Server Timeout: Specify the time the application will wait for an answer when making a request to a remote server. After that it will skip to the next url to process if no answer is received.

Default Cross Navigation Deep: Specify the default value for the deep value to navigate link from a listed web pages to other web pages (used when a new web pages are listed. (You can change this value)

Navigate also in page relative web address: allows to cross navigate in web pages expressed expressed as relative link (example: "../products/products.html") Url extractor will solve the link, will surf it and will process it.

Extract also relative web url: This setting is effective only if extracting web address. It will solve url in the form of relative web address (example: "../products/products.html") and will extract it as absolute (example: "<http://www.arpae.com/products/products.html>") saving it in the result.



Update

*** Update section is not available in the App Store Release (if you purchased via the App Store, to obtain an update use the App Store Update function)**

URL Extractor can inform you if an update is available.

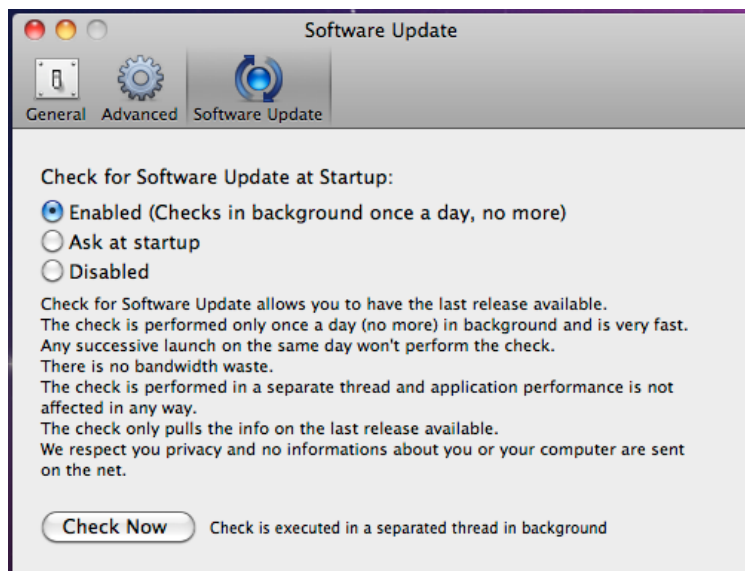
If enabled, the application will check no more than once a day.

'Ask at startup' asks you if you like to have the application check if an update is available when you start the application

'Disabled' avoids any check

When you check if an update is available, the application ping the Tension Software server and receive an answer about the last release available, the application can in that way know if its obsolete or not.

In case the application will ask you if you like to download an upgrade.



If you download the upgrade, you need to install it.

A common error is to download an upgrade and install it maintaining the old application somewhere on the hard disk.

Then using to open the documents sometime the old application and sometime the new one. This cause some problems*. To avoid it simply install your application in the place dedicated to it, the Application folder. When you install an update on the standard Application folder, the system will ask if you want to replace the old one with the new, answering yes will install the new one replacing the old.

*If the application says the document you are trying to open was created with a newer version of the application, probably you have two different release of the application on your Hard Disk and you are trying to open a document with the older version after having modified the document with the new one. When you receive a similar message open the about box inside the application and see if you are running the last release. In case download it, install it and USE it!

Help

URL Extractor provides a standard help menu ...maybe you already found it!

Under the help menu use the Visit Pomola.com to access our site a download the very last version of URL Extractor and other software for Mac we make.

Licensing the program

You can use the command under the Help menu to access our web site

From there you can purchase a license to use URL Extractor using one of the payment service we provide. It easy, fast, and secure.

Purchasing a license remove all the limitations inside URL Extractor

If you are evaluating to purchase this software, consider that it is realized in Cocoa and in Objective-C, the basis of all the last generation software for mac OS X, and you can expect a long life for any products based on Cocoa. Mac OS X and the Cocoa technology are estimated to have really really long life with continuous improvements (more then the life of Mac OS Classic of around 15 years) because Mac OS X and Cocoa are solid technologies (Mac OS classic was just a toy if confronted with Mac OS X) and new features can be added without problems as the time request them.

This software is Universal Binaries and runs native on both PPC Mac and Intel Mac.

If you decide to purchase a license, thank you for your support. When you see on the net how much quality software there is for Mac and how much software is added every day it is because users support small developers buying software.

Support

You can also obtain support using the 'Support Email...' command. An email will be prepared using your email client with the correct address to send to.

Yes, we answer to your emails.

URL Extractor is a Commercial Program

You can use our software for a test period of 10 days

After that you are required to buy a license to be legally authorized to continue to use our software

Licensing URL Extractor

*** In case of the App Store release a license is already included with your App Store purchase and you don't need to buy a license**

You can buy a license to use our software using the 'Buy License' command under the Help menu.

You will open in your browser our license web page on our web site

From there you can buy a license to use URL Extractor using one of the payment service we provide. It easy, fast, and secure and all most important form of payment are accepted.

We carefully selected or international reseller to be sure the buying experience for our customer will be as best as possible and without any problems.

After purchase you will receive from Tension Software a license email containing your full name and a license code

Open the License dialog with the command 'license' under the 'URL Extractor' menu and insert these data in the License dialog. The software will become fully licensed and fully enabled for future use.

Name and license code are remembered by the program and don't need to be re-inserted at successive launch.

In case you move on a new Mac you need to re-insert your name and license
The license is valid for a single Macintosh. You can purchase for two Mac as long as they are not used at the same time, as example a desktop and a laptop used by the same user.

Consideration about Licensing URL Extractor

*** In case of the App Store release a license is already included with your App Store purchase and you don't need to buy a license**

If you are evaluating to purchase this software, consider that it is realized using the Cocoa library (the native last generation library on Mac OS X) and in the Objective-C language (the first class language development on Mac OS X)

The two technology are the basis of all the last generation software for Mac OS X, and you can expect a long life for any products based on Cocoa and coded in Objective-C.

Mac OS X and the Cocoa technology are estimated to have a really long life with continuous improvements (more then the life of the now ancient Mac OS Classic which was around 15 years) because Mac OS X and Cocoa are solid technologies (Mac OS classic was just a toy if confronted with Mac OS X) and new features can be added without problems by Apple when they are required.

It was not so easy with previous generation of OS.

In our opinion the Cocoa technology will live and grow for many decades from now, so Cocoa based software are today the best way to go if you have to select software to use.

If you purchased a license, Thank you for your support.

When you see how much software there is for Mac and how much software is added every day, it is because users support small developers buying software and because the Macintosh software ecosystem it's really a great place to work due to the incredible success the Macintosh has today and the common feeling of the community (we use the Mac because we love doing things in a smart way).