



URL Extractor 4.8

User Guide



We Make Software - Ecleti.com

URL Extractor © 2006-2022 - Ecleti - Roberto Panetta all rights reserved

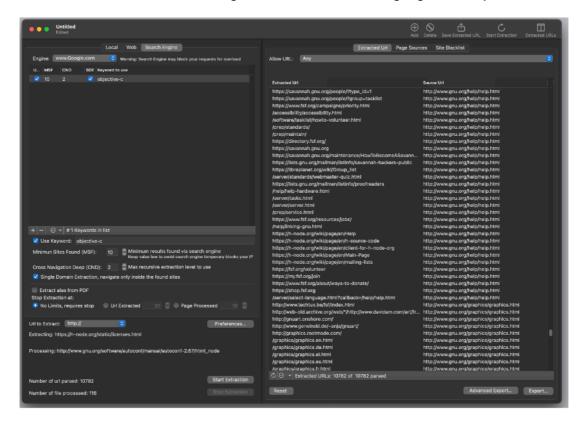
Every effort has been made to ensure that the information in this manual is accurate. Ecleti is not responsible for printing or clerical errors.

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 using search engines keywords. URL Extractor can work in the night, unattended, collecting big data for you.



Getting Started with URL Extractor

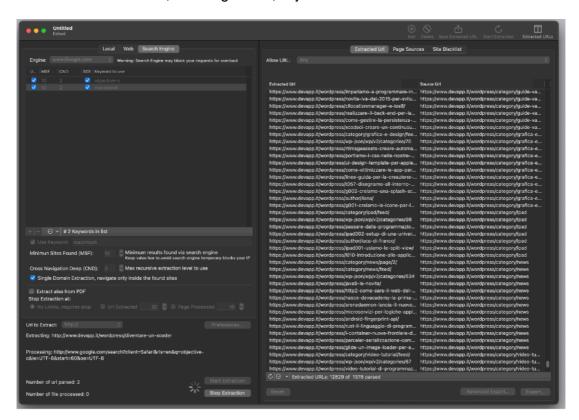
URL Extractor is absolutely simple to use

The interface is simple and intuitive.

The application is document based and you can save all the settings used in any particular extraction to a documents on disk.

The advantage is that you can repeat over time the same extraction (as example the same keyword search, once a week) without having to reinsert all the parameters. Just reopen the same document from disk and click 'Extract'.

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.



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

All this processing tasks (on disk and on the web) are executed in parallel, leaving your app responsive.

We use in this release 4, the latest parallel processing technology, all Cocoa native, using the incredible technology Apple made available on macOS.

The best use of URL Extractor is unattended, you launch the task and leave URL Extractor do the work for you. Working for enough time it can collect from the web a gigantic amount of URLs or emails!

Extracting from files and folders

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 a file or a folder (containing files) from disk.

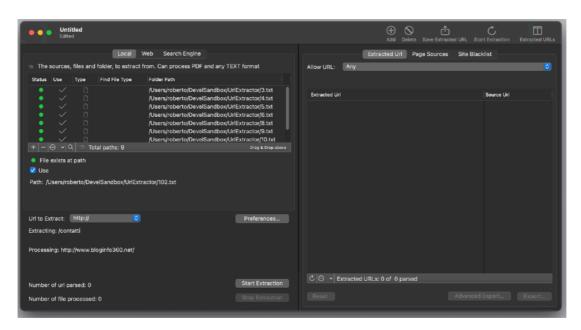
You can also drag and drop a file or a folder from the finder in the Url Extractor window

In case of a folder you need to specify which kind files you want to extract from (by extension)

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.

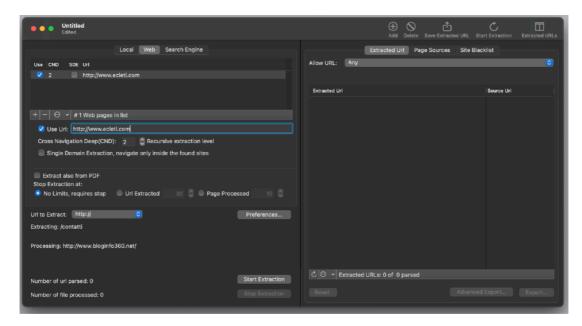


All the urls are now ready (on the right part of the window) to be saved on disk for you in text format.

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 see the url flow inside the 'Extracted URL' list on the right.

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

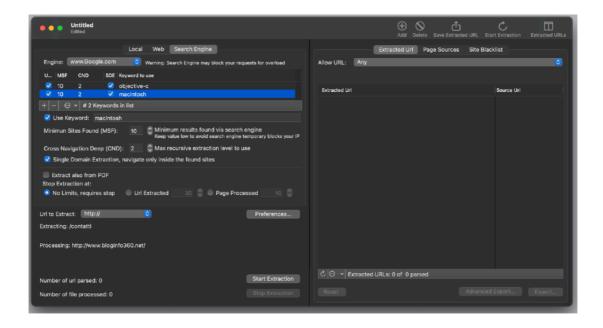
Extracting from search engine

Another way to extract from the web but starting from keywords (and not from web addresses) is extracting from 'Search Engine'

Select from the window the "Search Engine" tab

Use the '+' button and add some keywords 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.

The url are collected 'live' and you can see them fill the 'Extracted URL' in the right part of the window.

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.

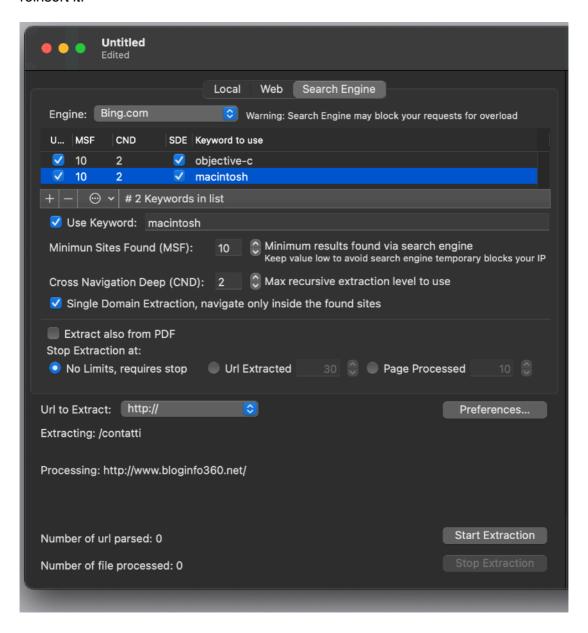
URL Extractor Reference

Application organization

URL Extractor is a standard Cocoa Document based application to extract URLs form your local disk or from the web using a series of user settings and data.

A window can be saved on disk as a URL Extractor document with all the settings and data 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.

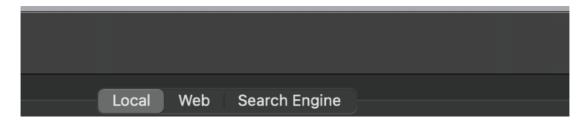


Extracting from local files, Web and search engine

Extraction operation can be performed on:

- •A list of file or folders (with all the contained files at any deep level) 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 "Local, "Web" or "Search Engine" you select which type of extraction to use.



In case of the Local selection of files and folder, all the files inside selected folders, also files inside nested folders at any level, will be processed to perform the extraction providing their extension is included in the "Find File types" field. They must be readable as text files or PDF 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. 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 "Search 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

Special note about PDF files

Text inside PDF files must be in in true text format to be processed and extracted If you have pictures of text they still are pictures, not true text so nothing can be extracted.

We decided to place this note because in another app we make named "To Text Converter" and available on the App Store, capable to convert PDF files to text, we get complaints about people trying to convert PDF of scanned pages (so just pictures embedded in pdf files) to text.

It can't work because there is nothing to translate to text, they are just pixels of a pictures, but they don't realize it.

They see a picture of text inside a pdf and can't recognize the difference between true text and a picture of text.

A picture of text is a picture, even if embedded inside a pdf file.

Here we have an example for you

The same text, on time as true text and the other as picture. Which is the true text?

Is this a text?

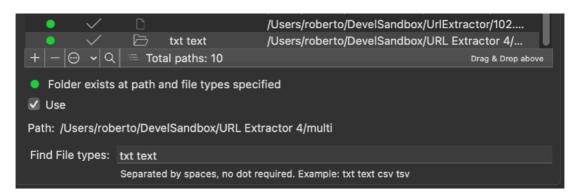
Is this a text?

One of the two is not a true text inside this PDF User Guide and could not be read by URL Extractor!

Which files to extract

If you select specific files 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.



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 or Search Engine 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.

You can also drag and drop any file or folder to the list area inside URL Extractor A list of path can NOT be imported because macOS now doesn't allow sandboxed app to access a file using as reference the file path string So it is the OS, not the app that introduces this limitation.

On modern macOS, with sandboxed applications, files can be accessed only if the user specify it via the selection dialog or via a drag and drop (for security reasons, to protect the user from malware or malicious software) unless they are resident inside the container (a dedicated folder for any app, usually hidden to the user, used for specific app files, not for user files)

Inserting a web page to navigate

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

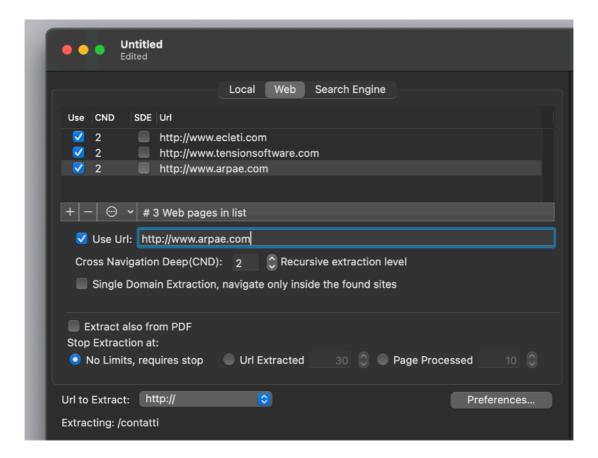
A list of web pages can also be imported form a file on disk

Web extraction

Url Extractor extracts from the web starting from a list of pages and for each one 'cross navigating' all the linked pages (jumping from one link to the successive) 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 URLs (email or others) from the web starting form a single web page!

To indicate the web pages to use to start the extraction process, you can edit them or just drag and drop them from other programs 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

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 going deep for 3 levels of link, 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.

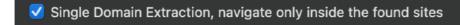
You can have web page that can be in the list and not used for extraction. Use the 'Use Url' check box to toggle between use and not use.



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 to continue)

Single domain extraction

Url Extractor can navigate only inside the file it's starting from (using the inserted url)



The default is to jump also outside, to linked sites. If you want to extract only from the starting site without jumping to other sites select 'Single Domain Extraction'. You will have a lot less result in that case.

In the other cases, when you disable 'Single Domain Extraction' the url represents a 'seed' from which URL Extractor just start to navigate, and the navigation, providing a high Cross Navigation Deep value, can be really unlimited.

To extract also from PDF (they will be recognized by the extension and read using the macOS native engine for pdf files) you must select the 'Extract also from PDF' checkbox



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 is required 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.

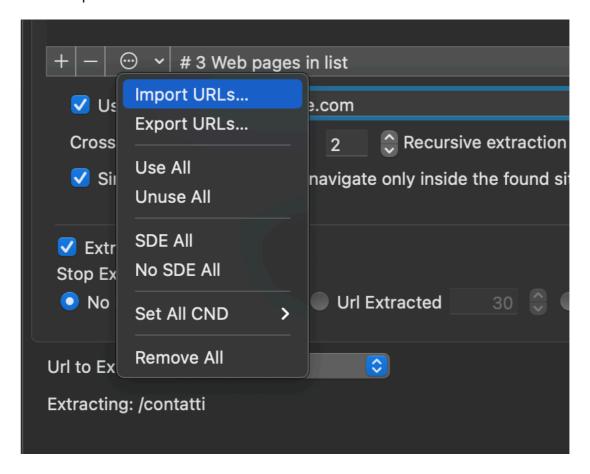
Page Processed: The program will stop to extract when the processed file number (web pages) reach the specified value.

Extracted URLs will flow inside the Extracted URL list.

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

Importing URLs to use

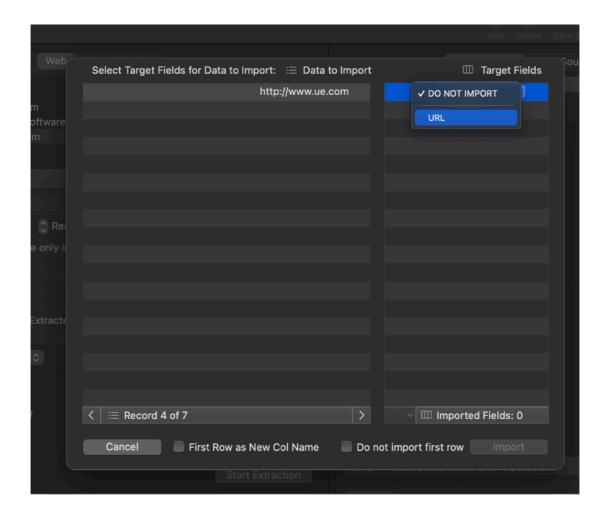
You can also insert a pre-existing list of URLs to be used for extraction, using the import command, from the tools popup at the base of the url list Select Import URLs...



A dialog will allow to select a file and after that which field to import as the url list to use

The panel allows you to browse the file you are importing to see it's content and select the right field to import in the application.

Select which field to import in the successive panel.



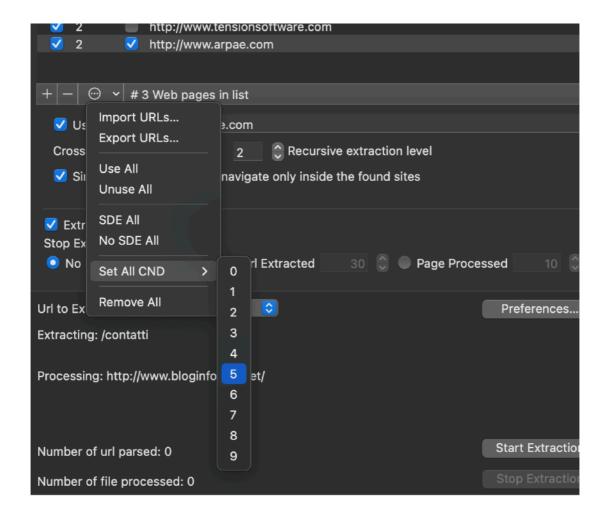
Then assign to the field the 'URL' popup value and click 'Import'

In case the file has just one column, indicate it as the source for the 'URL' destination.

Using the field at the bottom of the table you can assign the same 'Cross Navigation Deep' (CND) to all the imported url in a single command

You can also set to select all for use or as not-use

You can select all for SDE or not (Single Domain Extraction)



Exporting URLs used to extract from the web

Value used to extract 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.

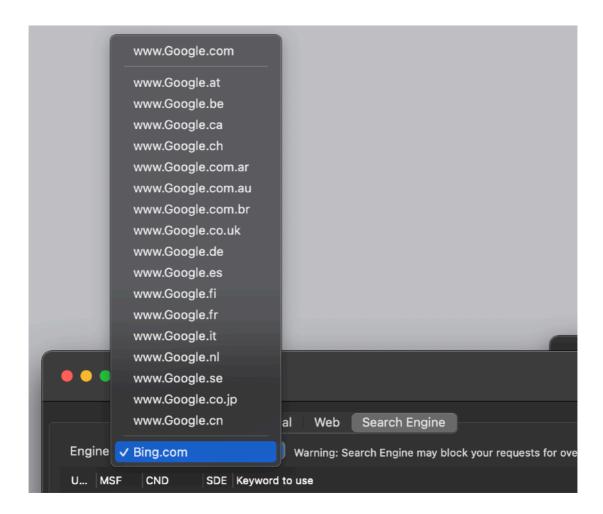
A single field (one column) file will be created with as many rows as the URLs to save

Search engine extraction

Search engine extraction is very similar to web extraction. It just adds a step (at the start of the steps chain) used to extract URLs.

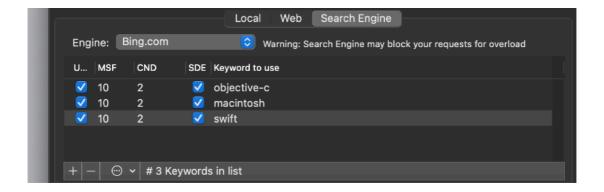
It searches on a search engine the resulting link associated with specified keywords, then it starts web extraction starting from all the links found and jumping to linked pages to continue its search.

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, adding editing and removing keywords, then launch the extraction.

URL Extractor will search (all unattended, without the need of human input) for any keyword the associated web pages and start a cross navigation extraction form any web page.

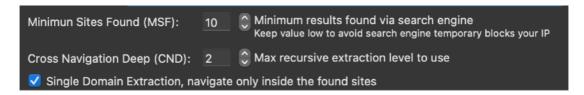


The software let analyze and extract a gigantic amount of data just starting from a single keyword.

It can do more with a list of keywords.

Selecting the corrects keywords 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 and the level of cross navigation deep to use for any associated site found on the search engine

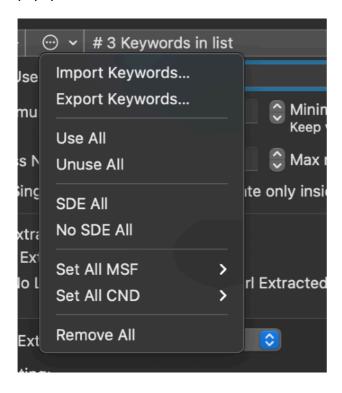


Url Extractor queries the search engine until the requested number of sites are found (associated to a given keyword), then it starts cross navigation and Urls extraction to the required deep.

Importing keywords

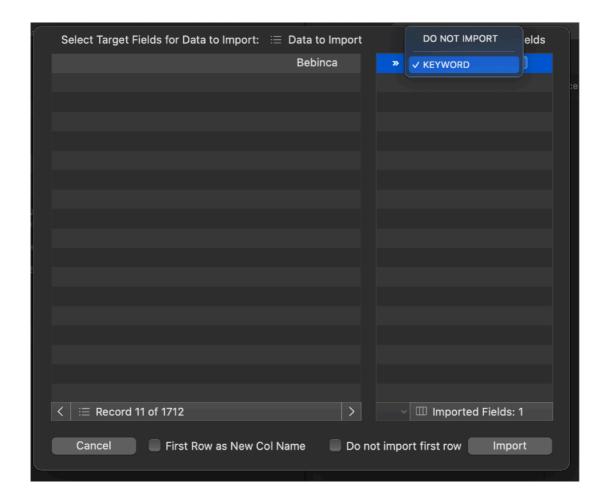
In a way similar to the URLs, also keywords can be imported from file on disk to the application

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



Select, in case the file to import has more than 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)

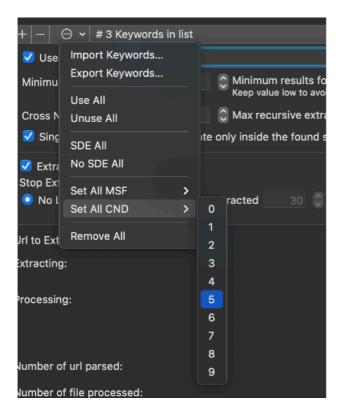


The panel allows you to browse the file you are importing to see it's content and select the right field to import in the application to 'Keyword' and leave a 'DO NOT IMPORT' for the other fields (if there are other fields in the text file you are importing from).

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 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

Url to extract

In any of the 3 method used, Local, Web and Search Engine, you can select wich 'kind' of url to extract.

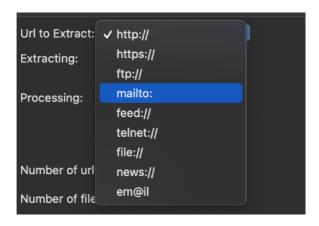
The Url to Extract popup menu is used to this purpose.

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

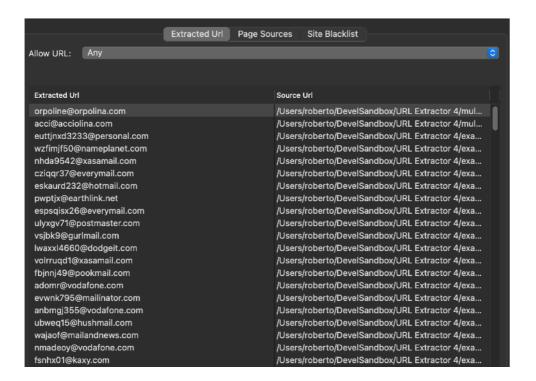
Pressing the "Start Extraction" button starts a extraction process. It is performed in a separate task leaving your application totally responsive

We employed the latest Apple technology to improve a lot the responsiveness of URL Extractor and we think we achieved a very good result over the previous release even if it was using a multithreading sophisticated technology.

In this latest URL Extractor release 4 we are using a new technology provided by Apple that is the suggested one for new apps that even surpasses (the now ancient) multithreading technology used in the previous releases and from our test results are really good!

Have a try yourself.

The extraction is saved in 2 columns:



In the first column are collected the 'Extracted Url' and in the second column is collected the url where the extraction comes from 'Source Url'

Usually the source url is a file references or a web address.

URLs 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://"

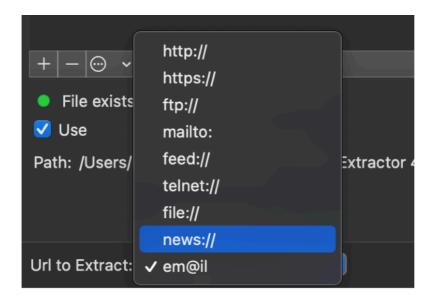
"telnet://"

"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.



Stop extraction

At any time you can stop Url Extractor and it will respond a fast way even if it is very busy with multi web requests over the web.

Here the latest Objective-C technology for multiple parallel tasks shows its best side.

In case you restart the extraction, it is restarted from the beginning 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

Extraction

The extracted URLs flow, as they are extracted, inside the 'Extracted URL' panel on the right part of the window

The 'Extracted URL' list can be resized using the mouse on the split bar. You can also hide and show the Extracted URL list via the toolbar or grabbing the split separator all to the right of the window

The user can put in place, in the 'Extracted URL' area, a filter to allows only some URL to be collected (or excluded).

To launch an extracted URL, double click it in the drawer or select it and use the load 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)

Export extraction as text

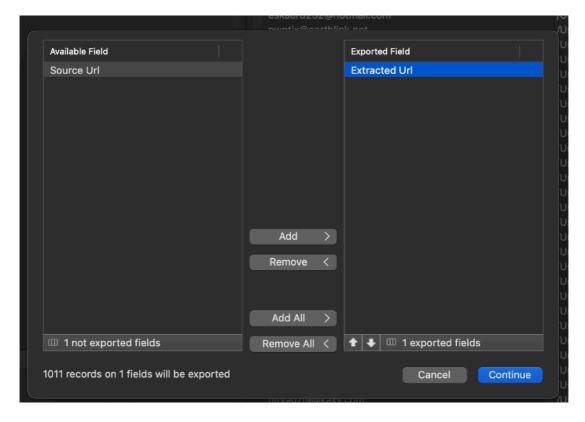
After an extraction all the url extracted are available in memory to be saved on disk. Pressing the "Export..." button you can choose where to save.

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)

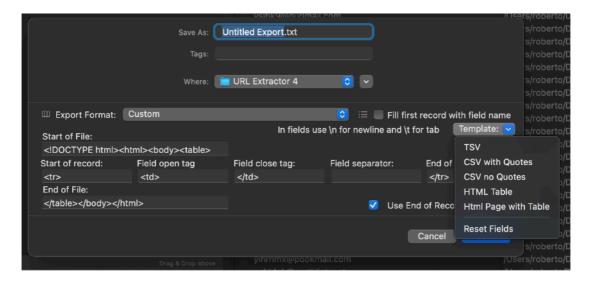


Pressing the 'Advanced Export' you can save both the extracted urls and/or the sources in 2 separate columns.



You can also choose to save in various format:

- Tab Separated Value
- Comma Separated Value with Quotes
- · Comma Separated Value no Quotes
- HTML Table
- Full page in HTML containing a table
- · Custom format where you can specify the separator you like



You can also select the option to insert the column names as the first row in the text file to export.

Later, importing the file, the first record can be used as column names the importing application can provide that

Please note that for free download version you must purchase a license to save the full extraction on disk.

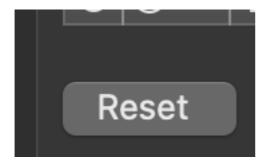
Unlicensed copy are limited just to evaluate the program.

Copies obtained from the App Store are already licensed and with full working capabilities.

Reset extracted data

The "Reset Data" button resets all the extracted data. After that no url are available in memory to save on disk as text.

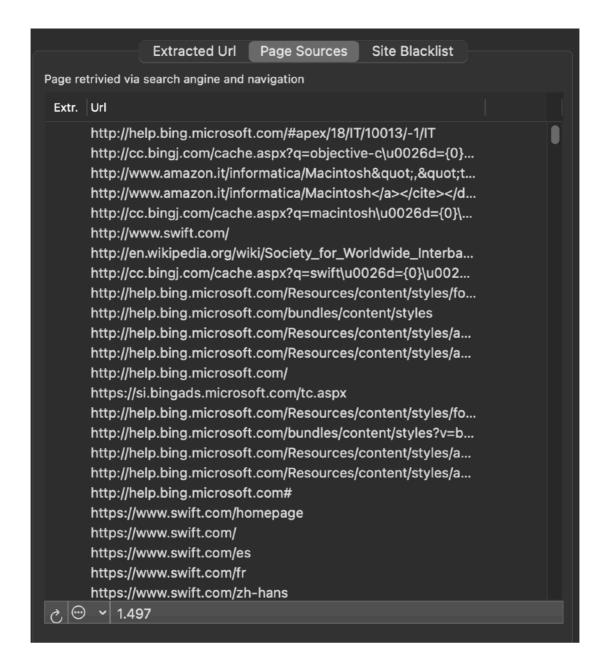
To have them again you have to extract it again using the "Start Extraction" button.



Page sources

In page sources the application show the pages it uses in both web and search engine extraction to extract the urls you requested

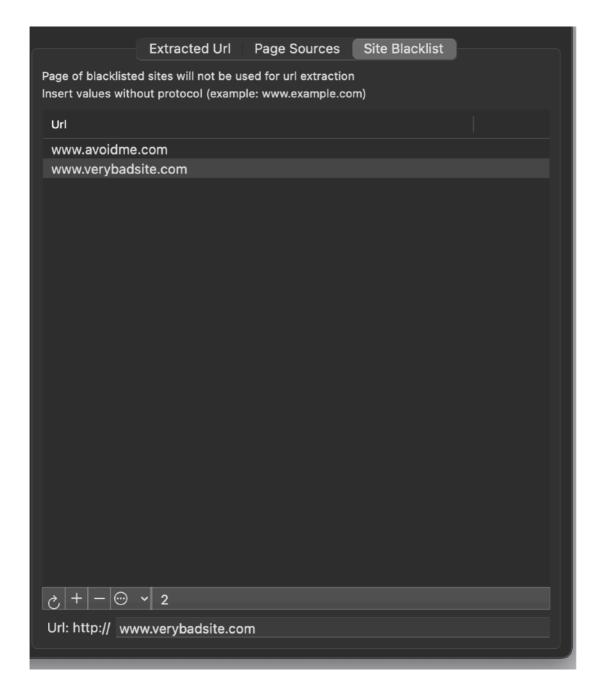
After a page is collected, it is visualised inside this list, and after it is being extracted it shows e green mark



It represent an indication of how the internal work to surf the web and extract it is going.

Site Blacklist

In Site Blacklist a user can define a series of domain that will not be used to navigate and extract urls



The user can add custom domain one by one pressing the '+' buttons or importing them from a text file using Import command under the popup menu at the bottom of the list.



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

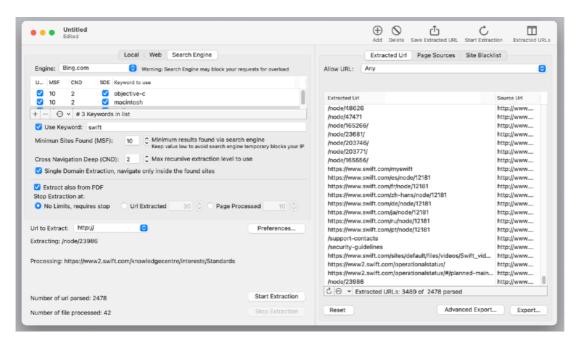
the total number of url extracted

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

The page collected and navigated inside the 'Page Sources' section. The url collected inside the 'Extracted Url' list

Dark Appearance Mode

This application is also fully compatible with Dark and Light Appearance Mode . You can switch at any time from one mode to another



General Preferences

General

Specify the action to do at startup

- Standard Behavior to use auto-resume and all the latest features of macOS
- 'Open Dialog if No Resume' to have the Open dialog opened if there are no files to auto-resume
- Open Dialog Always to have the open dialog opened at startup in any case

· Open Selected Document to have the selected document opened in any case



At any successive relaunch URL Extractor will execute the option selected. The additional 3 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
- Clear the selection

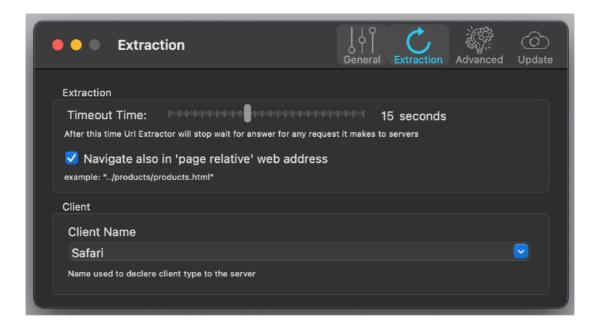
Extraction

Timeout Time:

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.

Client name:

the name used to declare to server which client application is requesting the page



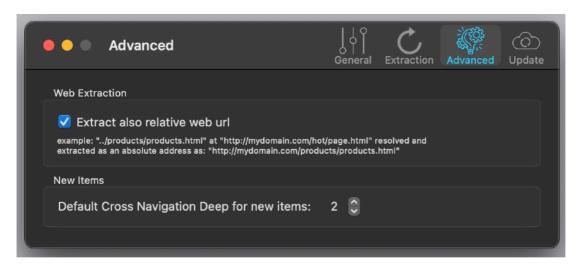
Advanced Preferences

Extract also relative web ur:l

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.

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)



Update

* This 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 notify in case an update is available It will inform you not more then once per day

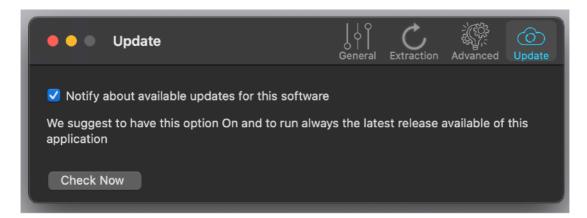
Search engines blocking you

After some queries via each engine you may notice URL Extractor may be not able to collect data from search engines anymore.

That may depend of search engine that may temporary lock out queries from your IP.

We have no control of it at all and it may depends of a lot of different factor, in any case the search engine may consider your queries a sort of 'spam' and temporary disable your requests.

In that case you need to stop sending any request the search engine and usually wait a bit before restarting your job.



Help

URL Extractor provides a standard user guide under the help menu ...considering you are reading it maybe you already found it!

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

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.

We support in the same way customer that purchased the software via the App Store or via other resellers.

URL Extractor is a Commercial Program

App Store apps are already licensed

If you downloaded an unlicensed version from the net, 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 reinserted 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.

About Tension Software and Ecleti

All the software asset by Tension Software is now under Ecleti New company name but still the same owner and main developer (Roberto Panetta) So for the users nothing changes. We provide the same software and the services. All the user licenses are valid and assistance is provided as before to all our users.