

# Webscavator User Guide



# User Guide Part 1: Adding, loading and editing cases

## Getting the web history files

---

Firstly, you must have a program that exports web history files into a text file format. The programs currently supported by Webscavator are:

- Fox Analysis - exports to CSV files
- Net Analysis - exports to CSV files
- Web Historian - exports to XML files
- Pasco - exports to CSV files
- Chrome Cache Viewer - exports to CSV files

You can add more programs by following the guide [here](#) (You'll have to do some Python programming though!).

### Using Fox Analysis with Webscavator

To obtain CSV data from Fox Analysis, go to File > Export To > CSV File. This will produce several files, please pick the one that ends with Website.csv.

### Using Net Analysis with Webscavator

To obtain CSV data from Net Analysis, click on File > Export History As > Tab Delimited Text. Rename the extension to .csv instead of .txt.

Sometimes Net Analysis produces files that do not conform exactly to Tab Delimited Text, and the converter will reject the file. If it is possible to locate the line(s) that cause the error, removing these will allow the file to be accepted.

### Using Web Historian with Webscavator

To obtain XML data from Web Historian DO NOT ask it to save as CSV. Web Historian currently uses a comma as a delimiter, and cannot be parsed correctly by any CSV parser when there are commas in the first field which is 'title' (very likely to occur). Instead save as an XML file.

### Using Pasco with Webscavator

To obtain CSV data from Pasco run:

```
> pasco index.dat > output.csv
```

### Using Chrome Cache Viewer with Webscavator

To obtain CSV data from Chrome Cache Viewer, select all the entries, and go to File > Save Selected Items. Save as a Tab-delimited Text File, and then change the extension from .txt to .csv. Make sure the file is in UTF-8

encoding, otherwise the CSV converter will not accept the file. You can do this by opening the CSV file in a program such as Notepad++, and choose Encoding > Encode in UTF-8.

## Starting Webscavator

---

To start Webscavator type the following command line in the Webscavator directory:

```
python launch.py runserver
```

In Windows, if you get the error:

```
'python' is not recognized as an internal or external command, operable  
program or batch file.
```

You will need to add python to your system environment %path% variable (this is easy to do, search online for *adding a path to system environment for [your operating system]*).

Finally, go to your favourite web browser (Firefox or Chrome is recommended) and go to <http://localhost:5000>.

## Adding a new case

Once you have got your extracted web history files, and started Webscavator, you are ready to make a new case. To add a new case, follow the 'new case' wizard.

### Add case details

#### Step 1: Add case information

Please give a name and filename for this case. All the data added for this case will be saved to an Sqlite database with the filename you have chosen. This means you can reload old cases and share cases between people.

Information

Case name:

File to save case to:

1. Type in the name of the case

2. Type the name of the file you want to save the case to. This must only consist of numbers and letters.

3. Click next

Case files are stored in the 'case files' folder

Each case is stored individually in its own Sqlite database in the 'case files' folder of Webscavator. You can store these somewhere else for safe keeping, but Webscavator will only load files from this folder. Having cases stored separately means you can easily share a particular case with a colleague without revealing any other cases.

## Add files

### Step 2: Add browser history files

Webscavator accepts CSV files produced from programs which process web browser files. You must process the raw web history files (e.g. index.dat files) first before uploading them to Webscavator. You can **add multiple different browser files** by clicking the "Add more" button.

Web Browser Files 1

Data name:  **1. Give a name for the data**

Data description (optional):  **2. Give an optional description**

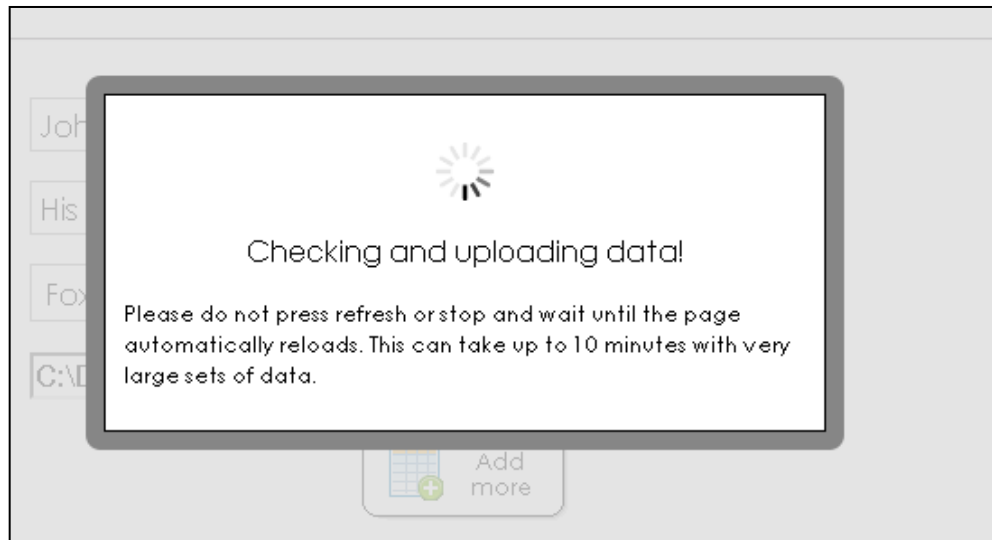
What program created this file?:  **3. Select the program you used to extract the web history files**

file:   **4. Upload the file**

**5. [optional] Add more data files. These can be made by a different program to this one**

**6. Click next (or previous if you want to change the case name)**

Make sure to select the correct program you used to make the file, otherwise the file will not be uploaded!



If the file is very large, it may take a long time to add all the data. You can add multiple files to a case if the person you are investigating has used multiple browsers or the program you have used has produced several files as output.

## Check details

### Step 3: Check data is correct

Case: Test

Date case added: 05:38PM 24 Aug 2010

File data saved to: test.db (SQLite database)

File location: C:\Documents and Settings\Sarah\My Documents\University\Thesis\Code\case files\test.db



#### Test web history:

File uploaded: webhistorian\_short.xml

Show  entries

Search:

Browser	Access Time	Type	URL
Chrome	13:58 10/06/2010	Auto Subframe - Redirect	http://docs.google.com/picker?relayUrl=http://www.google.co.uk/ig/images/rpc_relay.html&hl=en&hostId=gws&protocol=gadgets&action=loaded&title&minSize=800x600&icons=false&tpctoken=942944351&nav=([[photo-upload,from+my+computer],[photos,My+Ficora+Web+photos],[photos,Public+gallery,{type:featured}],([photos,Editor\,st+picks,{type:gwhome,mode:captions}]])&rpcUrl=http://www.ig.gmodules.com/gadgets/js/rpc.js?v=53e317d3b24e28cb3191a672d1dce87&container=ig&debug=0&c=1&view=[photos,Public+gallery,{type:featured}]&actionPane=legal&pli=1
Firefox	15:08 14/01/2010	Link	http://www.google.co.uk/search?hl=en-Gb&q=python+main&sourceid=navclient-ff&tz=1&GGGL_en-GbG8334G8334&ie=UTF-8
Firefox	15:08 14/05/2010	Link	http://www.google.co.uk/search?hl=en-Gb&q=python+main&sourceid=navclient-ff&tz=1&GGGL_en-GbG8334G8334&ie=UTF-8
Firefox	15:08 14/06/2010	Link	http://www.google.co.uk/search?hl=en-Gb&q=Python+main+way+to+test&sourceid=navclient-ff&tz=1&GGGL_en-GbG8334G8334&ie=UTF-8
Firefox	15:09 14/06/2010	Link	http://www.google.co.uk/search?hl=en-Gb&q=python+test+year+lots+of+words&sourceid=navclient-ff&tz=1&GGGL_en-GbG8334G8334&ie=UTF-8
Firefox	18:11 27/12/2009	Link	http://en-gb.start3.mozilla.com/firefox?client=firefox-a&ts=org.mozilla:en-Gb:official
Firefox	10:41 28/12/2009	Link	http://en-gb.start3.mozilla.com/firefox?client=firefox-a&ts=org.mozilla:en-Gb:official
Firefox	10:45 28/12/2009	Link	http://en-gb.start3.mozilla.com/firefox?client=firefox-a&ts=org.mozilla:en-Gb:official
Firefox	12:00 28/12/2009	Link	http://en-gb.start3.mozilla.com/firefox?client=firefox-a&ts=org.mozilla:en-Gb:official
Firefox	11:09 29/12/2009	Link	http://en-gb.start3.mozilla.com/firefox?client=firefox-a&ts=org.mozilla:en-Gb:official

Showing 1 to 10 of 1004 entries



4. Click next when you are happy with the data

1. Check the details are correct

2. You can sort by clicking on a column heading

3. You can see the next 10 entries by clicking on the arrow

## Finish wizard

### Case Added

Congratulations! The browser data has been added. [Click here to view your data.](#)

**1. Click on here to view the visualisations**

---

### MD5 Hashes


The MD5 hash for test.db is:

`b6affcf44fcb558b36e116b12dd557c0` **The hash for your case file**

All the hashes for your databases (with details such as time created and reason) can be found in:

`C:\Documents and Settings\Sarah\My Documents\University\Thesis\Code\case file hashes`

A new hash will be created if you edit any of the information by repeating the wizard or create a new filter.

[Previous](#) 

Congratulations, you have added a case. You can now see the visualisations.

## Security

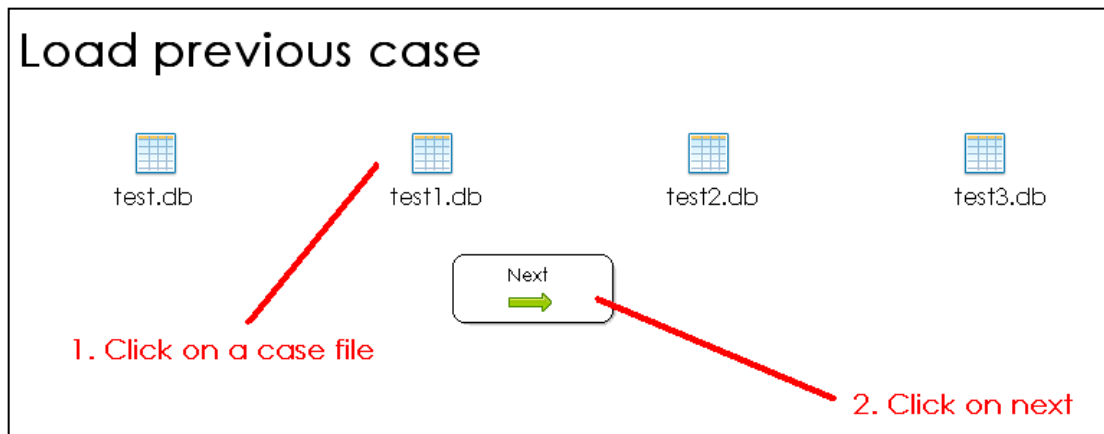
Once you have added, edited or loaded any case an MD5 hash of the Sqlite database is created and stored in 'case file hashes' folder of Webscavator, in a text file with the same name as the database file. For any subsequent edits, loads or filter additions, this file is updated with new hashes. The integrity of the database file can be checked by computing the MD5 hash of the database and comparing it to the latest entry in its MD5 hash file.

Webscavator has strong input validation. Because it only runs on a local port, the website is not available to anyone on the same network.



## Loading a case

---



If you have already added a case, you can load it into Webscavator. Make sure the case file is in the 'case files' folder of Webscavator, and then select it in the load case wizard.

## Editing a case

---

Once you have added a case or loaded a case, you can edit it by following the same wizard you did to create the case. You can change any of the details, including changing the files. To change the name of the case database, just rename the file in the 'case files' folder (but not whilst Webscavator is using it!).

# User Guide Part 2: Using the visualisations

## Overview

The screenshot below shows the homepage of Webscavator once data has been loaded in. You can add, edit or load a case using the buttons in the top right corner. Below the logo is a filter box with a set of prebuilt filters. Below that are the visualisation tabs, with a different visualisation on each. At the top of each tab on the right is a question icon. Clicking this will bring up information about that tab.

**WEBSCAVATOR**  
Filters for the Timeline, Websites visited and Online searches tabs

Home Edit case Load case New case Help About

Goes to this page (localhost:5000) Change/edit case

Filters

- Google searches
- Local Files
- Work hours
- Advert URLs
- Social Networking URLs
- Web Email
- News URLs

Reset Filters Add New Filter

Overview Timeline Websites visited Online Searches Files

Tabs with visualisations

Click to get help about this tab

### Overview for Testing

#### Heat Map of Internet Usage

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
00:00 - 00:59							
01:00 - 01:59							
02:00 - 02:59							
03:00 - 03:59							
04:00 - 04:59							
05:00 - 05:59							

#### Browser Statistics

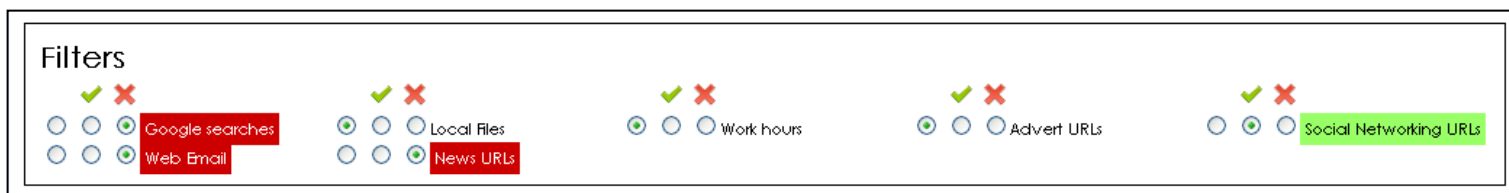
Firefox: 100.00%

#### Internet Usage Statistics

Average amount of web pages visited daily: 41  
Peak time of usage: 17:00 - 17:59

#### Case Information

## Filters



There are two kinds of filters available: those that remove/hide data from the visualisations and those that highlight/show data. Each filter has three radio buttons associated with it. Clicking on the button underneath the 'tick' will highlight data that corresponds to the filter, and the background colour of the filter will turn green. Clicking the button underneath the 'cross' will remove data that corresponds to the filter, and the background colour of the filter will turn red. The changes in background colour happen so that it is obvious that filters have been applied to the data, as it is not always easy at a glance to see which radio buttons are selected. All filters can be reset by clicking the 'Reset filters' link at the right hand side of the filters.

Removal filters always take preference over highlight filters. E.g. if a web history entry fits both a removal and highlight filter and they are both turned on, the entry will not be displayed as entries are removed first, then the remaining entries are highlighted.

### Adding a new filter

To add a new filter, click on the 'Add New Filter' link on the right of the filter box. This will bring up a pop up box. Enter a name for the filter and then you can choose what to filter on:

- Entry: can filter on the access date & time, the full URL, page title, HTTP headers, content type etc.
- Search Terms: can filter on the search engine used, search terms and term occurrence.
- URL parts: can filter on the different parts of the URL, e.g. the port, query strings, usernames, passwords etc.
- Web files: can filter on the program used to extract web history or different files uploaded. This is irrelevant if only one file was uploaded.
- Browser: can filter on browser name, version and profile used/IE location.

Some of the filters allow you to enter some text as the value, others you will select from a predefined list (e.g. 'browser name' will list all the browsers available to save you from typing), time values will ask you to select hours, minutes and seconds and date values have a date-picker (see screenshot on next page). The operation 'is in list' or 'not in list' will allow you to select text files located in the 'case lists' folder in Webscavator. These are simple files where each line is a value to compare to, e.g. a list of specific domain names you are interested in. You can add your own lists to this folder, and then select them in the select box. A few lists exist by default including a list of advert domains, social networking domains and online newspapers.

**Note:** If you are adding your own list, please make sure they are less than 450 lines as this is currently the maximum Webscavator allows.

Currently Webscavator does not support the editing and deletion of filters.

# Add a new filter

Filter Details

Filter quick name:

Create Filters

Entry

- Pick...
- Full URL
- Page title
- Modified Date
- Deleted
- Modified Time
- File Name
- Access Date
- Content Type
- Directory
- HTTP Headers
- Type
- Access Time

Submit

*select data type*

*select an attribute. A selection of operations to perform will then appear.*

# Add a new filter

Filter Details

Filter quick name:

Create Filters

Entry  Contains  AND

Browser  Is  AND

Entry  Greater than  :  :  AND

Entry  Is  AND

Submit

*Filter name*

*Add multiple filter lines*

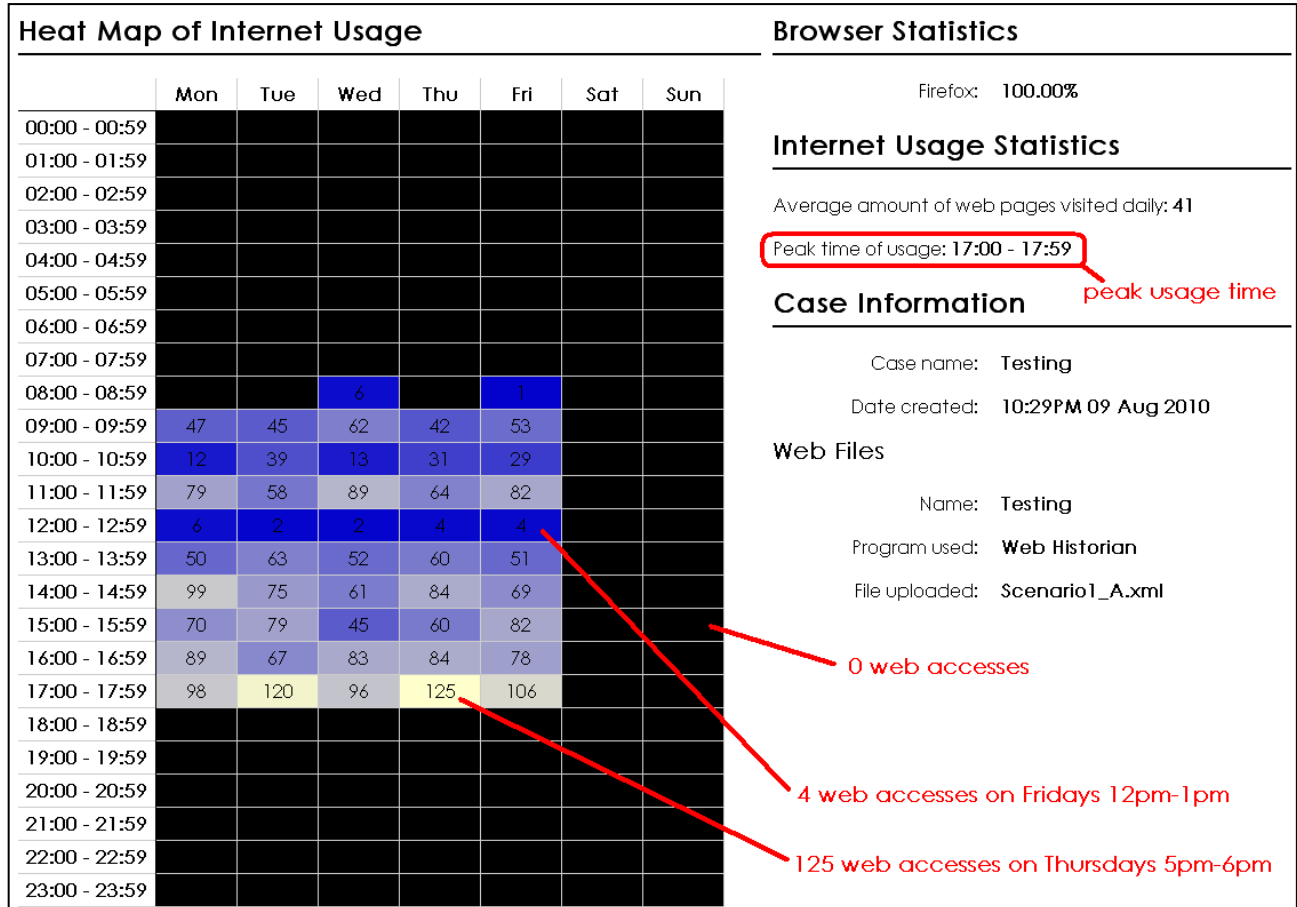
*Delete this filter line*

*Submit your filter*

Sun	Mon	Tue	Wed	Thu	Fri	Sat
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	1	2	3
4	5	6	7	8	9	10

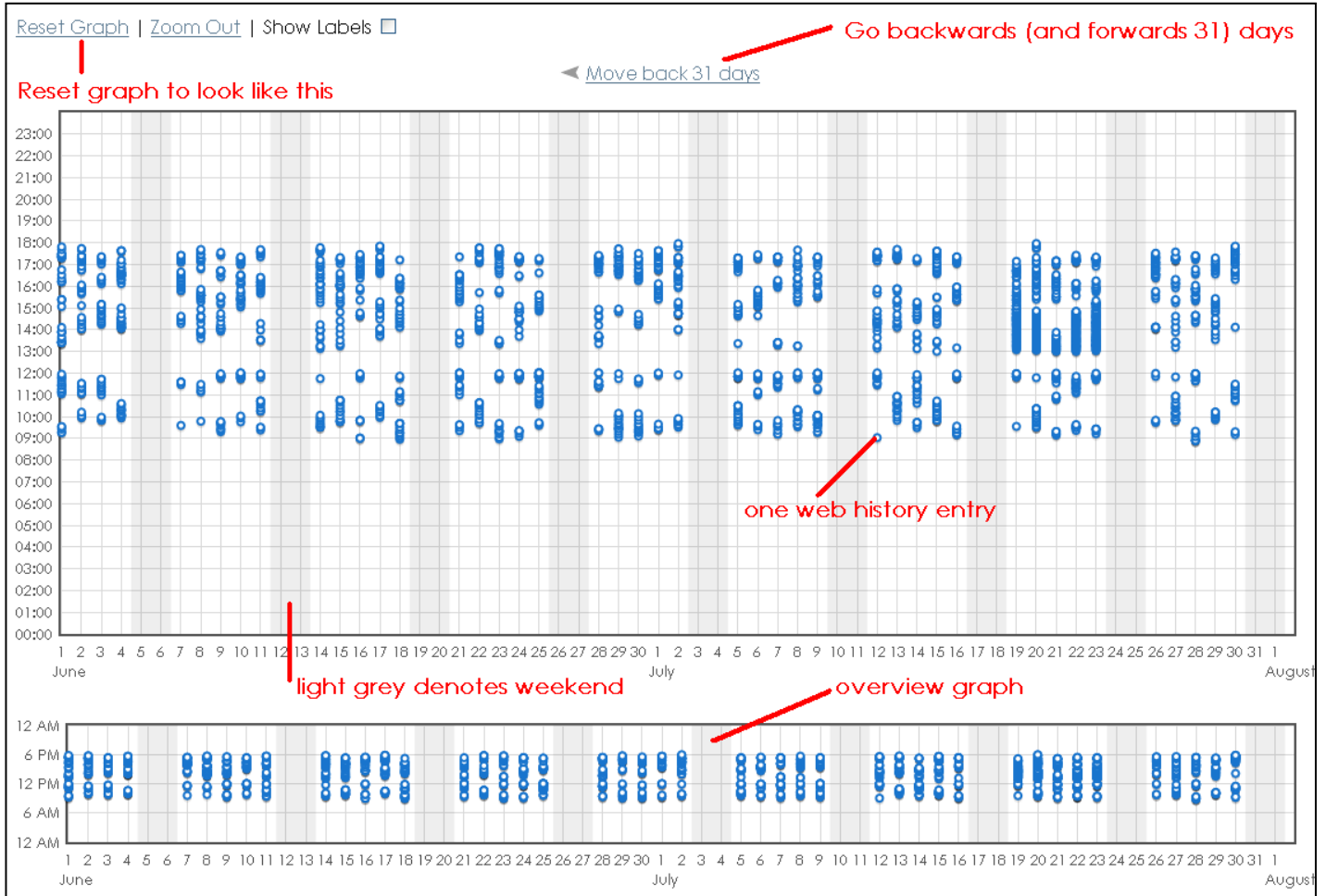
## Overview Tab

The overview tab has a heat-map showing the amalgamated number of entries for each day of the week for each hour. The more entries, the lighter the colour. Black means there are no entries. To the right are some overview statistics and information about the current case.

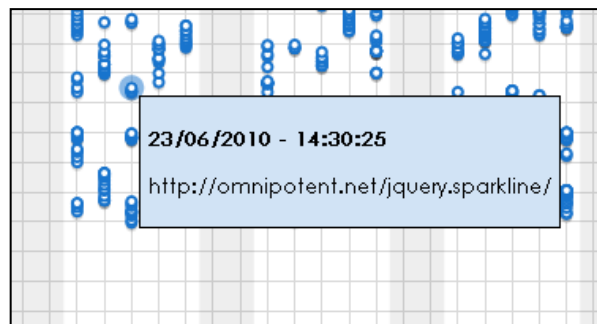


## Timeline Tab

The timeline shows the web history in two graphs. Along the x-axis is the time, and the date along the y-axis. Each point is a web history entry. The smaller, bottom graph is an overview graph, and does not zoom in or change colours like the larger graph. This is to keep a sense of context. To move about, click and drag the top graph around, or click "move back 31 days".



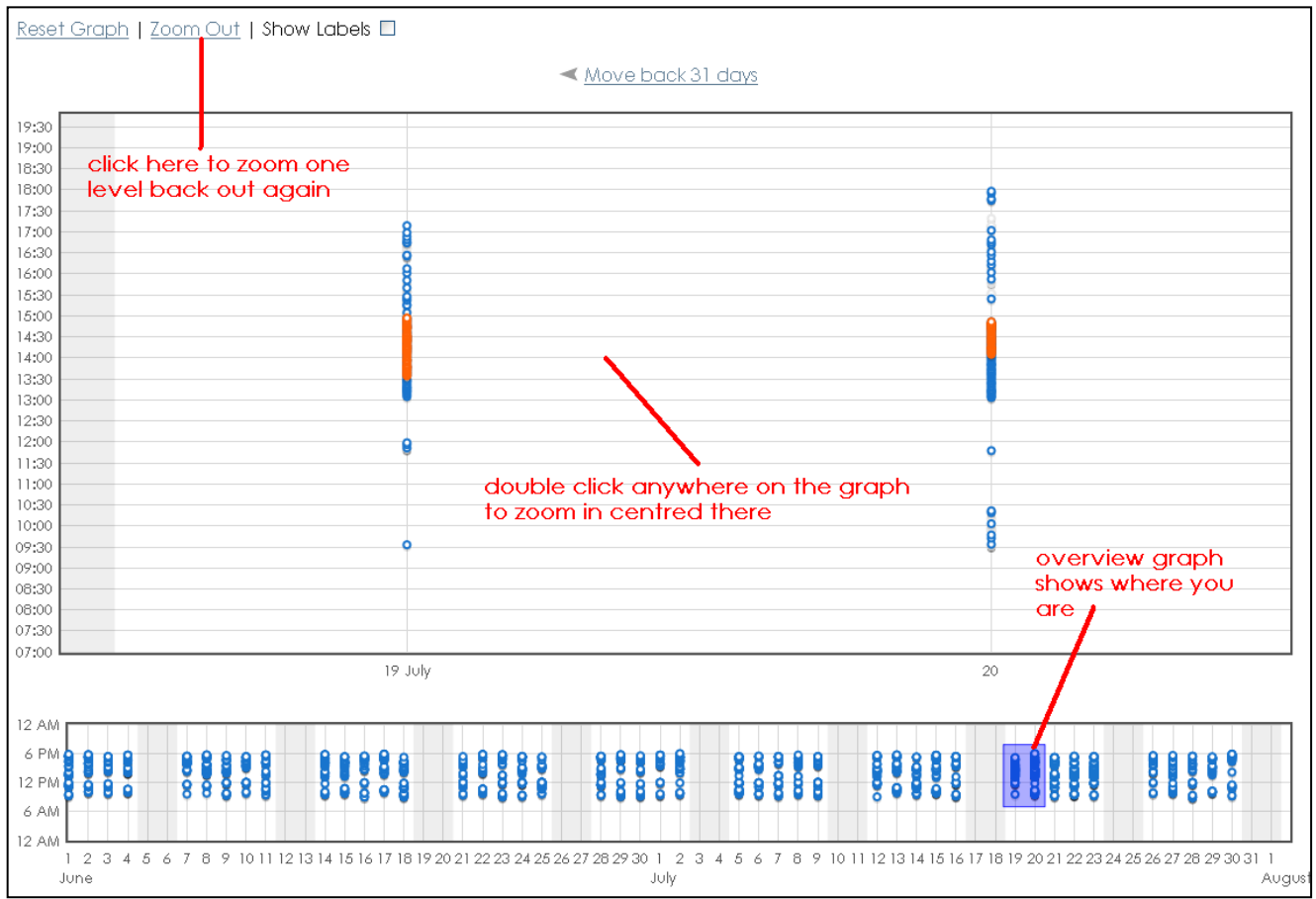
Hovering over a point gives the date, time and URL. Clicking on a point will produce more information below the graphs such as browser, page title etc.



Applying a removal filter will make any points that correspond to that filter turn a very light grey. Points are not completely removed from sight to keep a sense of context. Applying a highlight filter will make any points that correspond to that filter turn orange.



There are two ways to zoom in. To zoom in slowly, you can double click on the top graph. This will zoom in centred where you clicked. Otherwise, you can drag a blue rectangle over the overview graph, and the top graph will zoom in there. The overview graph will always display a blue rectangle showing where you are zoomed into. You can still pan around by dragging your mouse. Clicking on 'Zoom out' will zoom out a small amount. Clicking on 'Reset graph' will zoom out completely.



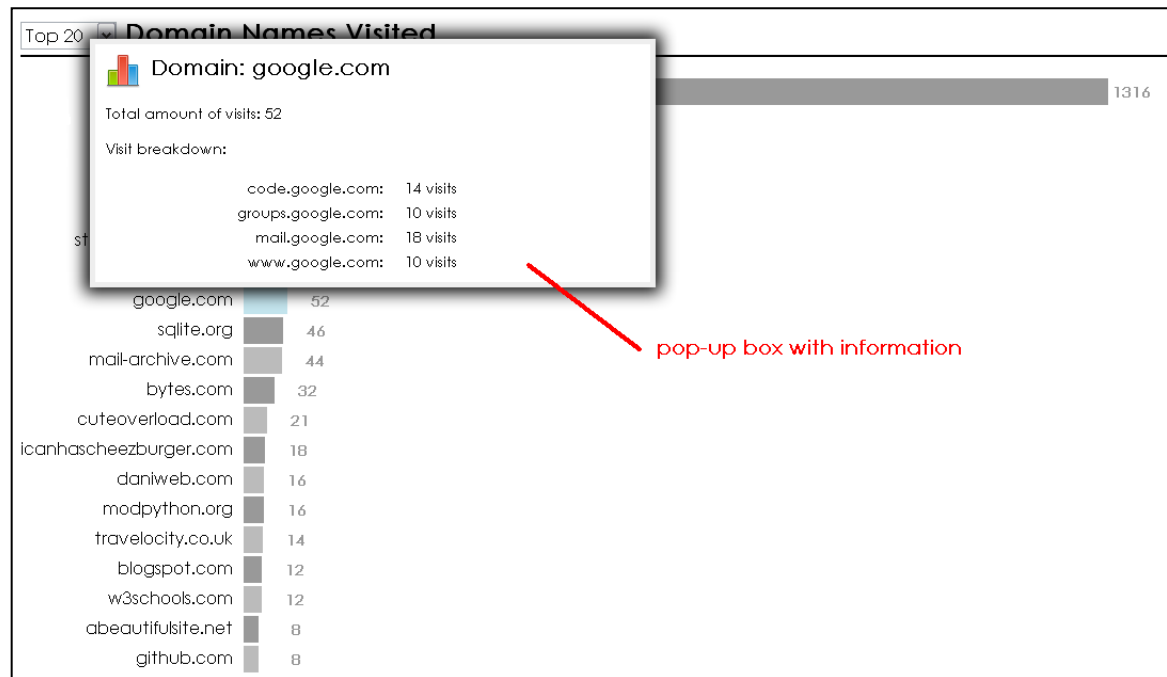
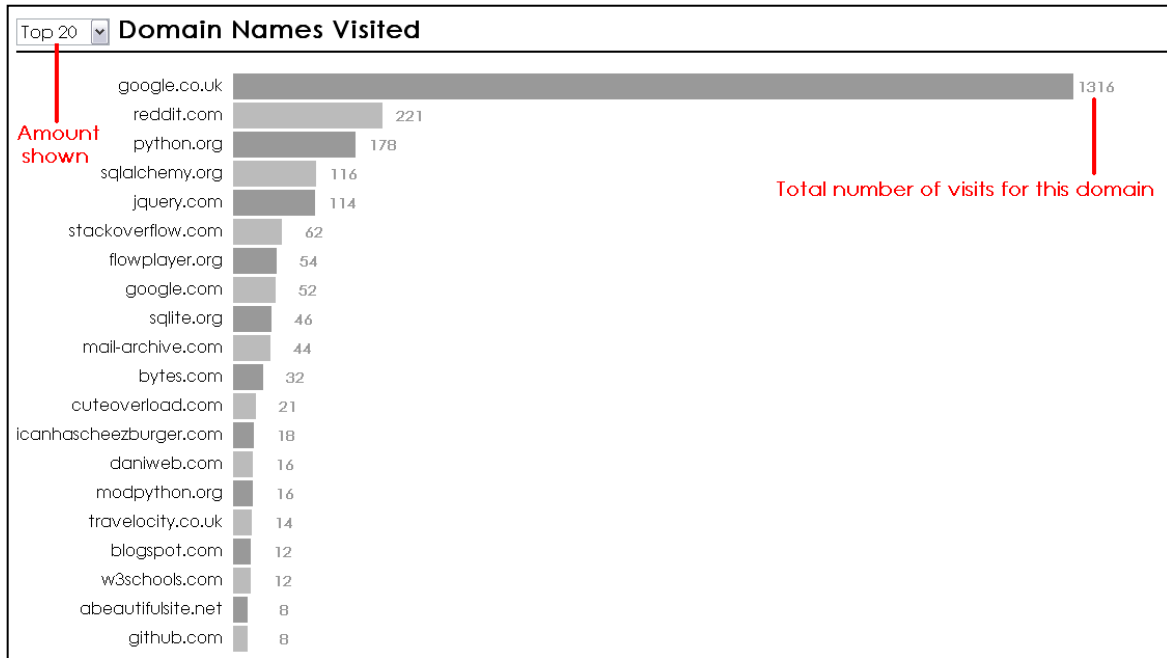
When you are zoomed in, you can check 'Show labels' which will show the URL beside the point.





## Websites visited Tab

This tab has a bar chart of the top domain names visited. You can change the amount viewed from Top 20 to Top 50, Top 100 or All. Clicking on a bar will bring up a pop-up with more information. The bar chart will not show data that corresponds to any removal filters that are active. If any highlight filters are active, only the data corresponding to those will be shown.

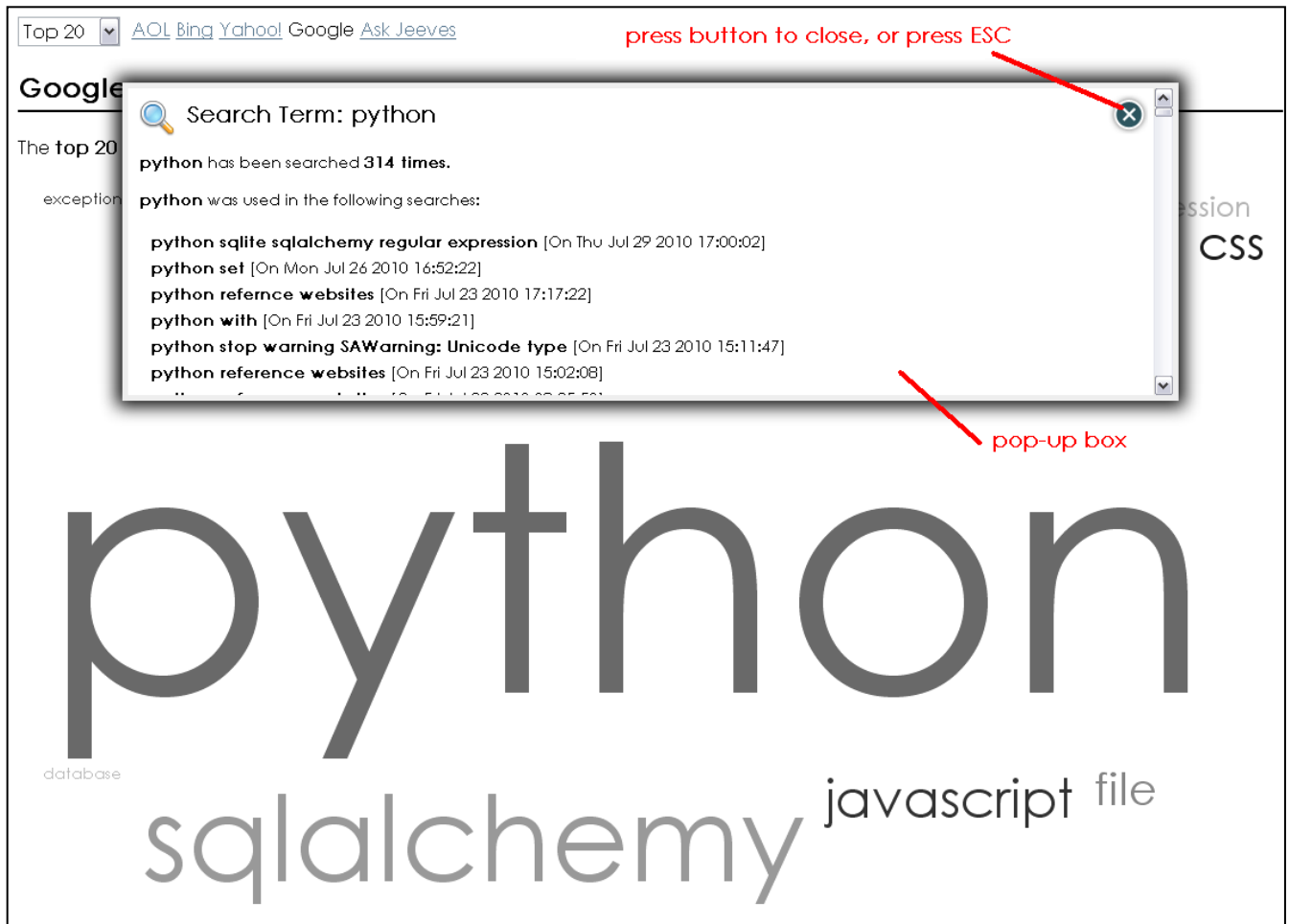


## Online Searches Tab

This tab shows the online search terms queried using search engines. Click on a search engine name to see the search terms queried. You can change the amount viewed from Top 20 to Top 50, Top 100 or All. The word cloud will not show data that corresponds to any removal filters that are active. If any highlight filters are active, only the data corresponding to those will be shown.

The larger the word, the more it has been searched for. Clicking on a word will show a pop-up with more details such as all the searches this word appeared in with the date and time.

The screenshot displays the 'Online Searches Tab' interface. At the top, there is a dropdown menu set to 'Top 20' and a list of search engines: AOL, Bing, Yahoo!, Google, Ask, and Jeeves. Below this, the 'Google' search engine is selected. Two red arrows point from the text 'Amount shown' and 'Choose the search engine' to the dropdown and search engine list respectively. The main content area shows 'The top 20 searched terms are below.' followed by a word cloud. The largest word is 'python'. Other prominent words include 'jquery', 'sqlite', 'string', 'css', 'sqlalchemy', 'javascript', 'file', 'exception', 'connect', 'get', 'list', 'datetime', 'regular', 'tools', 'remove', 'time', 'one', and 'expression'. A red arrow points from the text 'Search terms. The larger the term, the more it has been searched.' to the word 'jquery'.



## Adding new search engines

It is possible to add more search engines by editing the config file located in the webscavator/config folder. Let's assume a typical search with the new search engine is `http://search.foo.com/?terms=search+terms`, and this search engine is called "Foo Search". All Foo Searches contain 'search.foo' in the URL, and 'terms' is always followed by the search terms. Add an entry underneath the [search\_engines] and [search] titles like so:

```
[search_engines]
search.foo = terms

[search]
search.foo = Foo Search
```

# Files Tab

This tab shows all the local file accesses divided up by drive. You can click 'show details' to show a pie chart of the file types on this drive, followed by a breakdown of the files accessed. Click on a file to pop-up details about when the file was accessed.

Currently only Windows drives are supported.

**Files Accessed for Testing** ?

In total, 205 files were accessed 406 times on 3 drives.

**H:/ Drive [6 accesses]** [show details](#)

---

**C:/ Drive [385 accesses]** [show details](#)

---

**D:/ Drive [15 accesses]** [show details](#)

click on show details to see the files accessed



**D:/ Drive [15 accesses]** [hide details](#)

Pie Chart of the different file types accessed

**File Types for Drive D:/**

■ Images  
■ Documents

**Images [12 accesses]**  
**Documents [3 accesses]**

D:  
 images  
 stuff  
 names of security guards.docx [1 access]  
 how to make the fake id card.docx [1 access]  
 jewels in price order.docx [1 access]

Pie chart of different file types



click to show files



Files shown a file directory format



