SiteOrbiter

Concepts and User Manual

Introduction

SiteOrbiter is a tool designed to aid the web developer and administrator in web site management. The software provides visualization of link relationships and provides HTML and comma delimited reports. SiteOrbiter can be used to quickly find and debug broken or moved links, or used to look at HTTP headers. In addition, it can provide PDF output of the particular map to facilitate documentation of a particular site.

This document serves as a reference and usage guide to SiteOrbiter.

Please note that SiteOrbiter is compatible with Apple OS 10.1 and upwards, and cannot be installed on Apple OS-9 or earlier operating systems.

SiteOrbiter is not guaranteed to load all sites. Sites with large amounts of embedded content, such as media formats (i.e. Flash) or dynamic JavaScript may not load correctly or be detected.

Installation

Download the disk image from http://siteorbiter.cc/download.html. Once completed, double click on the downloaded file icon. It will open to display the SiteOrbiter application. Drag and drop this application icon in your Applications folder, home directory or other desired location. The disk image Finder window can be closed and the disk image "ejected". SiteOrbiter is now installed and ready for use.

Registration

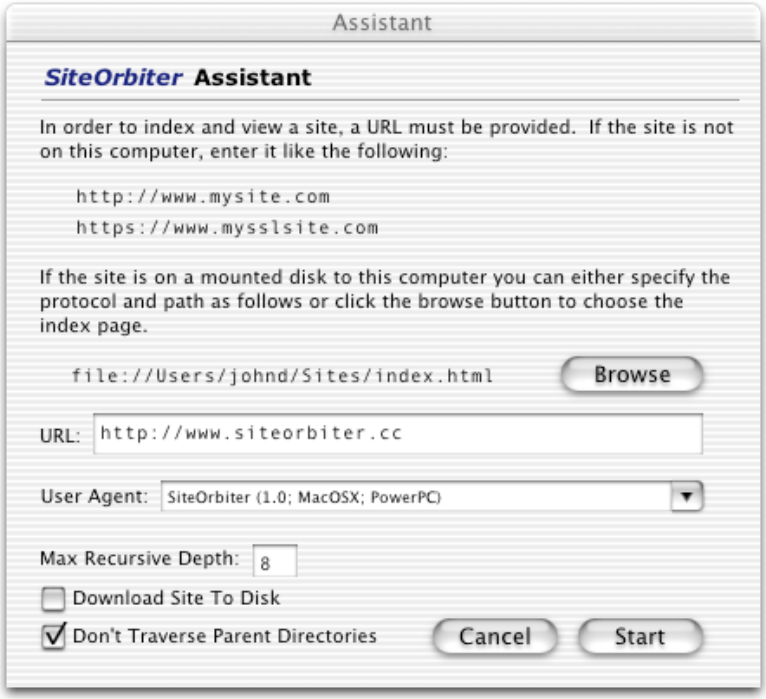
SiteOrbiter can be registered at siteorbiter.cc for a fee of $20 dollars. This fee helps support the development of SiteOrbiter and makes it possible to bring this software to the Macintosh. The

end user has a 20 day free trial period, after which SiteOrbiter will start prompting you to register the product.

Indexing a Site

When SiteOrbiter starts, the program will display an Assistant to aid in the process of mapping and gathering details about a particular web site. There are several options, each of which is explained below.

Figure 1: SiteOrbiter Assistant



Enter the URL of the site you wish to map, or click browse to find a site on your local computer. The supported protocols are HTTP, HTTPS and a mounted file system. Optionally, if you exclude the protocol, http will be assumed.

Select the User Agent string SiteOrbiter will present to the web server. The User Agent string specifies the type of browser making the request. SiteOrbiter has several of the main browser strings by default, or you can specify your own by directly typing into the User Agent box. This is ignored if the protocol is file.

> ➢ Some sites are designed to dynamically deliver data based on the browser doing the request. If you find that a site doesn't respond to the SiteOrbiter user agent, you might try a more standard browser setting, such as Mozilla.

Set the maximum recursive depth. This indicates how "deep" SiteOrbiter will descend into the mapped site. The default is 8, enough to get most small to medium scale sites completely. Any pages beyond the maximum recursive depth are indexed, but not visited. Note that if you have a lot of unvisited links after the mapping process, the depth setting may need to be adjusted upward to completely index the site.

By default, SiteOrbiter doesn't traverse parent directories. This means that if you enter a subdirectory in the URL address box, SiteOrbiter will not ascend to the parent links, and thus not gather information "up the tree". This is useful when you wish to map only a section of a web site.

To initiate the process, click "Start". SiteOrbiter will then start to map the specified site. On large sites, this process takes awhile.

Working with Previously Saved Documents

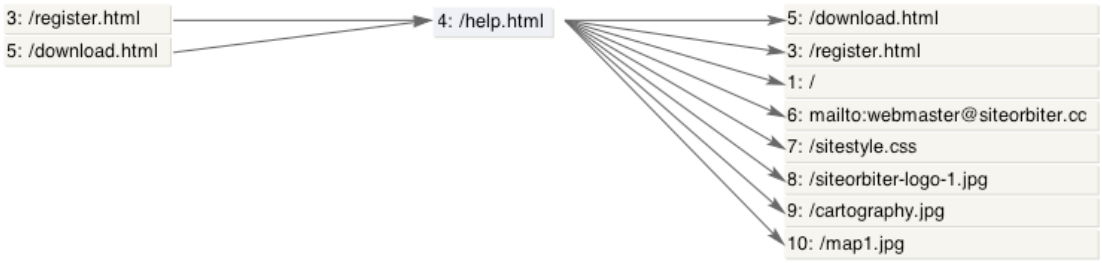
To load a document instead of indexing a new site, click cancel. The Assistant will close, and you can open a previously saved site from the File open menu. Be aware that a site with thousands of links can take a while to load as SiteOrbiter works out the dependencies.

Documents are saved with the extension `.stor`, which identifies to the operating system that this is a SiteOrbiter formatted file. These are plain text files and can be modified using any standard text editor. The `.stor` files are different than the .txt extract format (see The Site Data Table section).

The Map Window

Once SiteOrbiter has finished loading a document, or indexing a Site, it will display a map of the Site. Depending on processor speed and number of elements, the map can take awhile to display. There are two viewing modes, "All Objects" and "Subset". Subset is the default. The view mode can be switched back and forth via the Display Mode selector in the lower right hand corner of the document.

Figure 2: The Subset Map Display



> ➢ To navigate in subset mode, click on a link text in the map window. It will become the selected node, and will subsequently be displayed with all inbound and outbound links.

Sometimes, it's not necessary to view an entire website as a complete map, like in the All Objects mode. To make it easier to understand link relationships to a selected node, the Subset view mode can be selected. Use the Display Mode selector in the lower right hand corner of the document window to accomplish this. To navigate in subset mode, click on a link text on the map window, and it will become the selected node, and will subsequently be displayed with all inbound and outbound links.

Figure 3: The Display Mode Selector

To view all the objects on the map, select "All Objects".
To enter subset view mode, select "Subset".

Figure 4: The All Objects Map Display – A Simple Site

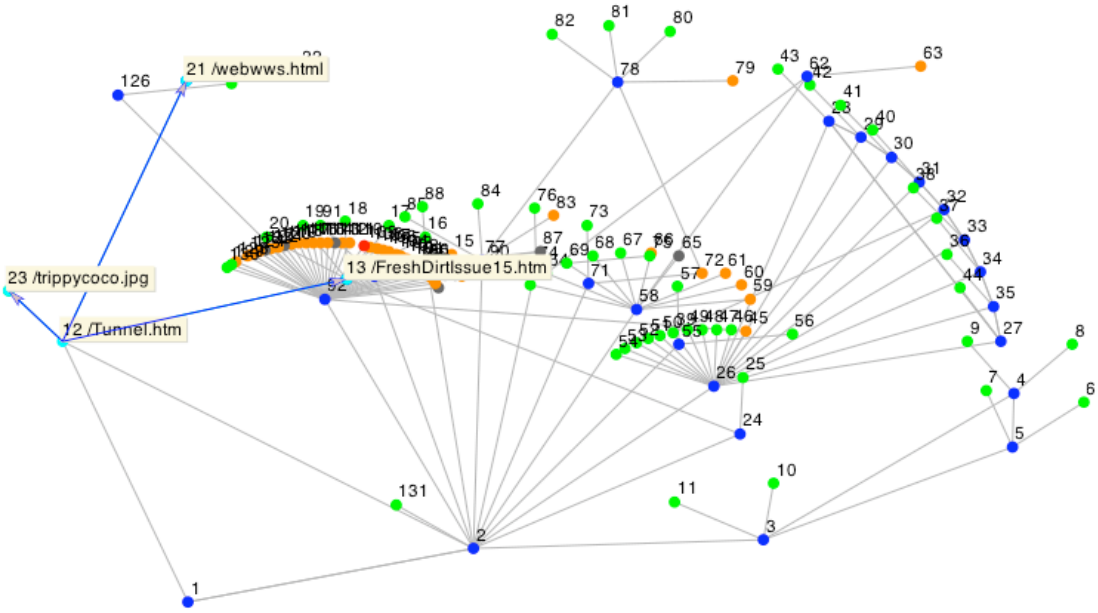
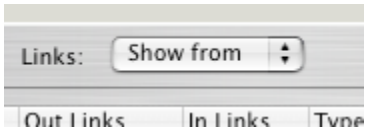


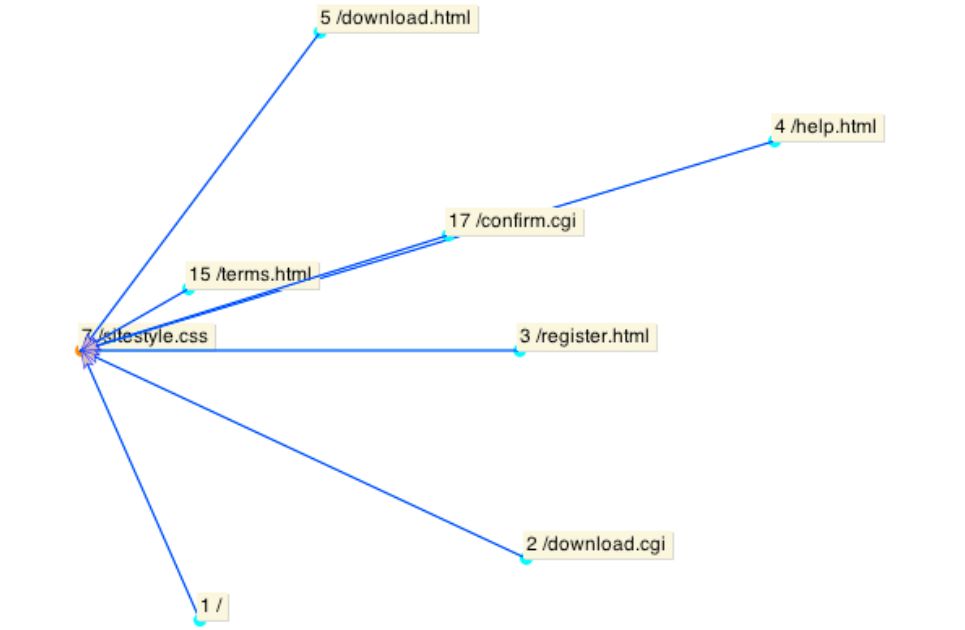
Figure 4 illustrates a map of a relatively uncomplicated site. The starting page is always identified as 1 (the first element). If the node is clicked via the mouse, it will be highlighted in both the map and the table. In addition to the node itself highlighted, the outwardly connecting links from the node will also be highlighted. To view all pages that connect to the highlighted node in question, use the drop down box labeled Links:

Figure 5: Link Direction Selector (All Objects Mode)



The "**Show from**" mode shows all outbound links from the selected node. The "**Show to**" mode highlights all nodes that link to the selected node. For example, to identify all nodes that rely on a particular cascading style sheet, highlight the style sheet node, and select the "Show to" mode.

Figure 6: All nodes that link to the selected node – All Objects Mode



To help with organization, links can be moved by dragging the node to a new location on the map.

Links between nodes are either drawn as black or grey. Black indicates SiteOrbiter encountered this node for the first time within the site. Grey indicates that the node was in cache, and already known to SiteOrbiter.

Nodes are color coded based on two factors: page type and return code. The color key indicates the color encoding:

Node Color	Meaning
Blue	Content type: text/html and was found (200) ok.
Green	Content type is an image, and was found OK.
Orange	An external link that was found and available.
Purple	Unknown (unreachable or different protocol such as a mailto: link)
Red	Page not found. 404
Grey	Page returned, but no content type was indicated by the server or detected by SiteOrbiter.

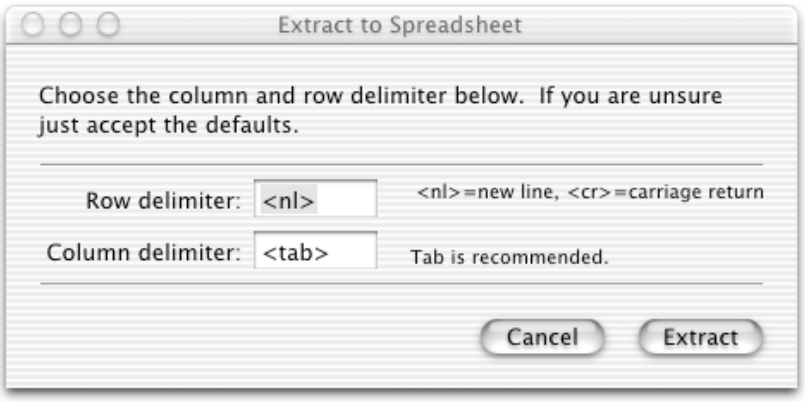
Turquoise	The currently highlighted node.

The Site Data Table

The data table shows you at a glance some of the characteristics of the highlighted nodes. Clicking on a row on the table will highlight the node on the map, and conversely, clicking on a node will highlight the node in the table.

Table data can be exported via the Extract button, or the File menu. The defaults are tab delimited, one row per line but this can be modified in the Extract To Table Dialog.

Figure 7: Extract to table



The delimiters can be any character, but be careful, as some delimiters (such as commas) can appear in the data set. The data can then be loaded into any spreadsheet system that accepts imported data.

The data set can be searched through via the Find command, or Meta – F. SiteOrbiter will scan the site characteristics and report any matching data in any column. It will highlight the node and the corresponding row in the table, and place in the messages window the string that caused the match. Looking for the same string again will cause SiteOrbiter to start at the last place that data was found.

The Document Controls

The document controls help the user to manipulate the map or change some view characteristic to gain a better or more particular presentation.

Figure 8: The Document Controls

Hide Map: On large documents redrawing the view can be computationally intense, so the view can be turned off to improve performance. If checked, the Hide Map switch will cause the view to disappear.

Copy Site to Disk: If this is on, the site will be downloaded to disk when it is re-indexed. This is useful for staying up to date to a remote site. Note that all site data will be downloaded, not just the differences.

The Shift Keys: They move the entire map left, right, up or down in the map view.

The Only Show buttons: These buttons filter by either the currently highlighted link, or only allow the site to display HTML pages, suppressing the drawing of image nodes or other data.

Zoom: This slider causes the map to scale in or out, allowing you to see the site at different scales. The **0** button resets it to it's standard scale.

Show To/From: The "**Show From"** mode shows all outbound links from the selected node. The "**Show to**" mode highlights all nodes that link to the selected node.

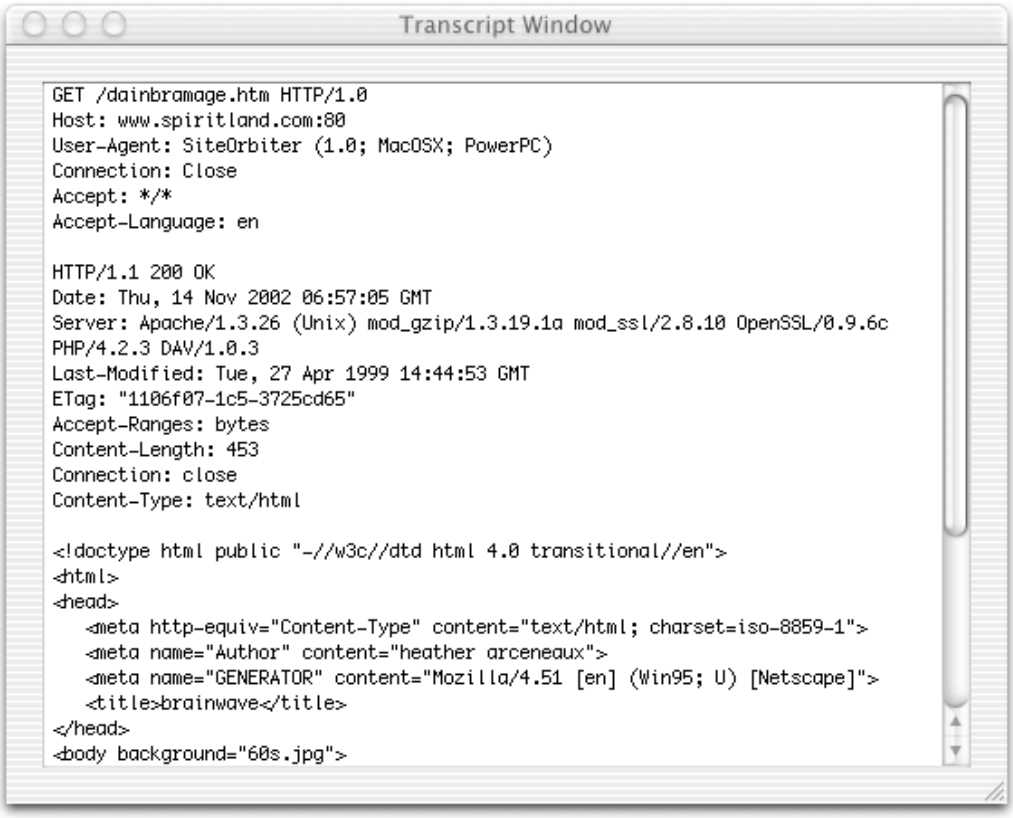
Locked: This causes the highlighted node to stay highlighted, even if other nodes are clicked on. It is useful for organizing a subset of nodes on the map without loosing the highlight.

Show Names: If on, names will be drawn next to there respective nodes. This can cause the map to look very cluttered, and is turned off by default.

Transcript Window

The transcript window allows you to gather header information. This can be very useful when debugging CGI or dynamic content pages. It also allows you to identify what cookies a page may request to be set.

Figure 9: The Transcript Window



Viewing the robots.txt file for a site

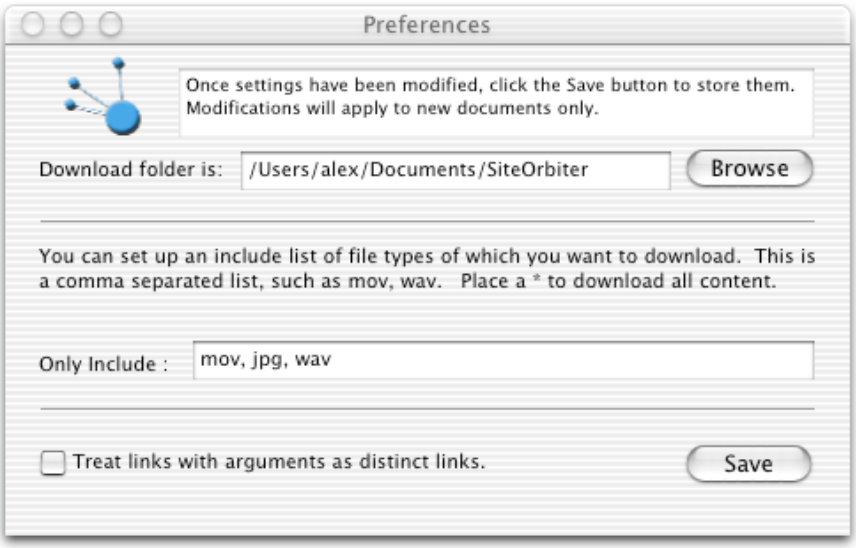
Many sites have defined a robots.txt file which is located in the root directory of the target site. This file defines where a web crawler ("robot") such as SiteOrbiter may not traverse. SiteOrbiter, when connecting to a remote site, first checks to see if there is a `robots.txt` file available. If there is, it will abide by it. This feature cannot be turned off in this release. If a robots.txt is defined, the "View robots.txt" text in the lower left hand corner of the site window will be enabled. Click on this to bring up a transcript window containing the contents of the `robots.txt` file.

If a site fails to load, check to see if the robots.txt file prohibits it with a /.

Preferences Window

The preferences window modifies the global behavior of SiteOrbiter. These behaviors include where to download web content to, what types of web content to include, and whether or not to include links with arguments as distinct in the map.

Figure 10: The Preferences Window



To revert to the defaults, just clear the entry that you wish to restore, and click on the Save button. The program will populate any blank entry with the default value.

Be aware that treating links with arguments as distinct means that a page could be counted multiple times. For example:

www.mymappedsite.com/program.cgi?argument1 and
www.mymappedsite.com/program.cgi?argument2

will map to separate nodes on the map. This is probably not the behavior you want, as the map will appear to contain many more links than there actually are. This can be useful if testing a CGI based content delivery system.

The View Menu

The **View** menu provides access to several different reports. It is a convenient way to get information on a particular node or characteristic of the map. For example, to produce a report on all the missing and not found links for the indexed site, use **View->Missing and Moved Links**. This will start your default browser and produce an HTML report.

The **Links By Directory** presents a table of directories and each link within the directory.

The **Last Index Report** retrieves the report generated by the last re-index operation. This is useful to view changes that have occurred between indexing operations.

The **Links To/From Selected** field produces an HTML page that contains a table of all pages that link to the selected node, and all pages that the selected node links to.