Robots Exclusion Protocol

The Robots Exclusion Protocol is the protocol for instructing search engines whether or not to index, archive or summarize any given page. These instructions are contained in a robots.txt file in the root (or other) folder of the site. The robots.txt definitions are advise to robots, they do not actually block access to data. Well-behaved robots will obey many of these directives.

Disallowing features

Many TikiWiki features are provided through specific programs, so crawling those features can be stopped by blocking the specific programs. If you are using Search Engine Friendly URLs (SEFURL) those should also be listed. You should consider the characteristics of the search site and your content; for example, if you have images with short descriptions then those are not of much use to search engines which do not handle images well.

Disallow: /tiki-calendar.php Disallow: /tiki-browse_categories.php Disallow: /tiki-browse_freetags.php

Blocking duplicate access paths

Wildcards are new features which many robots do not yet recognize. Some robots will recognize some wildcards. "*" means any characters. "?" means the question mark. "&" is an ampersand. "$" represents the end of the line.
If your site is not using SEFURL, many parts of the site have to be accessed with at least one parameter, such as "?id=123", thus you should not block access to many patterns with a question mark in them. If you are using SEFURL then your URLs will have fewer question marks.
For TikiWiki, "Disallow: /*&" could be used to disallow every URL with an ampersand, which will avoid having robots trying to examine variations of pages. However, you should examine your site to consider whether you want to block all URLs with an ampersand or only specific parameters. Some default URLs require at least one ampersand, such as accessing a file's information requires specification of both a gallery ID and a file ID. By adding a parameter after the ampersand you can disallow specific parameters.
For example, if a robot fully crawls a file directory in the default order there is no need for it to also follow the URL with "&sort_mode" and view the same data in a different order.
It is obvious that the early 2009 version of the Cuil crawler robot, twiceler, crawls every identified variant of a URL. It is not known whether twiceler obeys the * wildcard.

Disallow: /*&fullscreen= Disallow: /*&popup= Disallow: /*&sort_mode=

Keywords:
security

Robots.txt Directives

DIRECTIVE	IMPACT	USE CASES
User-agent: *	Says following section is for a specific robot. Asterisk is for all robots.	Different settings for specific robots. Each named robot then ignores contents of "*" section, so repetition is often needed.
Crawl-Delay: 1.0	Asks robot to delay the specified number of seconds between queries.	Slow rate of robot crawling. Some robots ignore this. Googlebot ignores it and Google webmaster site gives some control.
Disallow	Tells a crawler not to index your site — your site's robots.txt file still needs to be crawled to find this directive, however disallowed pages will not be crawled	'No Crawl' page from a site. This directive in the default syntax prevents specific path(s) of a site from being crawled.
Allow	Tells a crawler the specific pages on your site you want indexed so you can use this in combination with Disallow	This is useful in particular in conjunction with Disallow clauses, where a large section of a site is disallowed except for a small section within it
$ Wildcard Support	Tells a crawler to match everything from the end of a URL — large number of directories without specifying specific pages	'No Crawl' files with specific patterns, for example, files with certain filetypes that always have a certain extension, say pdf
* Wildcard Support	Tells a crawler to match a sequence of characters	'No Crawl' URLs with certain patterns, for example, disallow URLs with session ids or other extraneous parameters
Sitemaps Location	Tells a crawler where it can find your Sitemaps	Point to other locations where feeds exist to help crawlers find URLs on a site

HTML META Directives

DIRECTIVE	IMPACT	USE CASES
NOINDEX META Tag	Tells a crawler not to index a given page	Don't index the page. This allows pages that are crawled to be kept out of the index.
NOFOLLOW META Tag	Tells a crawler not to follow a link to other content on a given page	Prevent publicly writeable areas to be abused by spammers looking for link credit. By using NOFOLLOW you let the robot know that you are discounting all outgoing links from this page.
NOSNIPPET META Tag	Tells a crawler not to display snippets in the search results for a given page	Present no snippet for the page on Search Results
NOARCHIVE META Tag	Tells a search engine not to show a "cached" link for a given page	Do not make available to users a copy of the page from the Search Engine cache
NOODP META Tag	Tells a crawler not to use a title and snippet from the Open Directory Project for a given page	Do not use the ODP (Open Directory Project) title and snippet for this page

Alias names for this page

Robot | robots.txt | Robots