

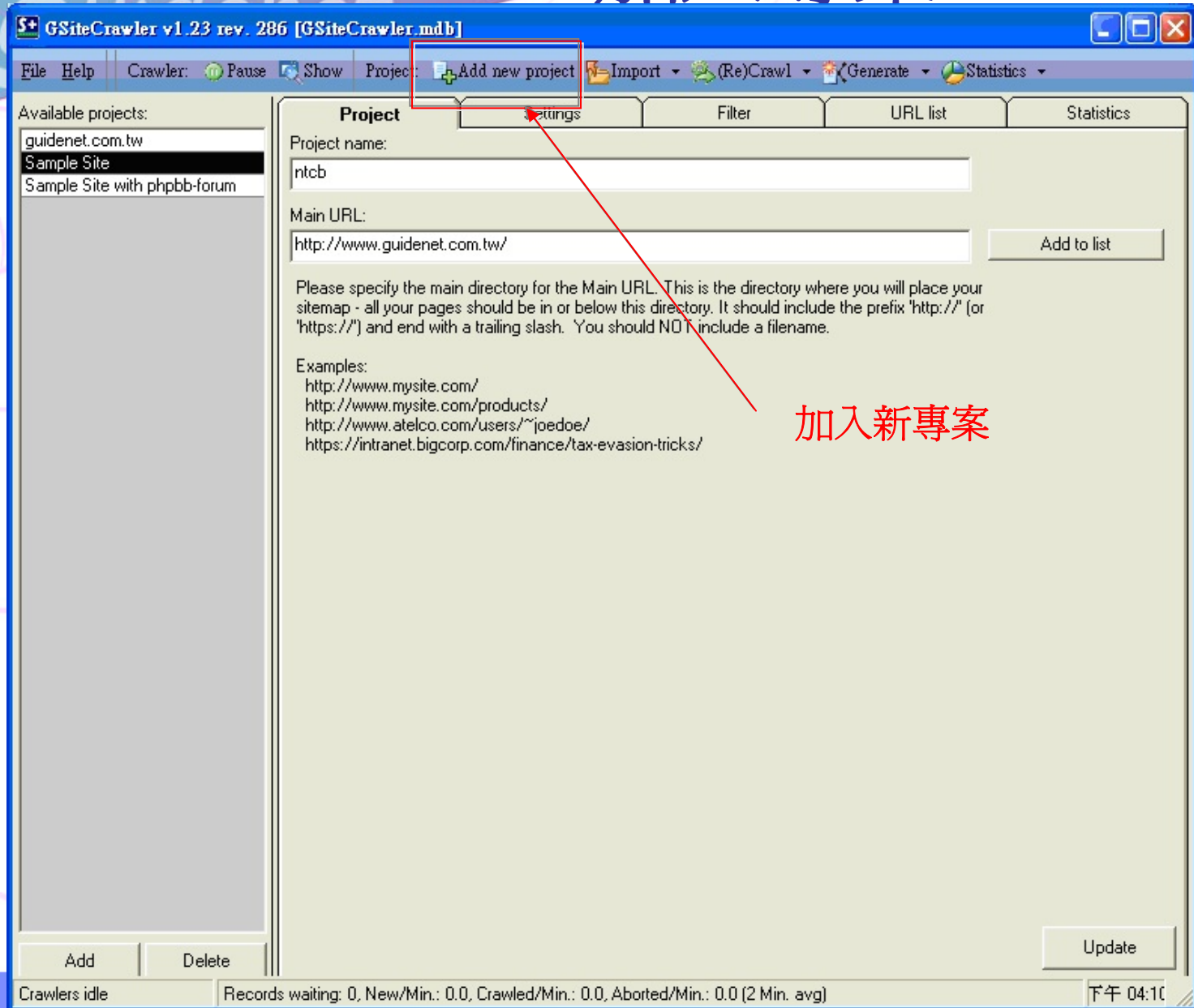
SEO與網站優化與網路行銷與搜尋引擎與Webometrics專家

GSiteCrawler 使用說明

 知諸網科技股份有限公司

<http://www.guidenet.com.tw>

加入專案



加入新專案

加入專案

Guidenet

GSiteCrawler - New Site Wizard

Welcome to the GSiteCrawler files

This wizard will help you get a GSiteCrawler files possible.

To use this wizard, you will only need a few facts you can even use this to check what you wanted to :-).

If you have FTP-access and want to start right away, you will - of course - need an account.

Step 1 of 4: Website address

Please enter the address (URL) of the website you wish to generate a sitemap file for here. A project name will be suggested, although you are free to change it.

Main address of your website (directory):

Please specify the main directory for the Main URL. This is the directory where you will place your sitemap - all your pages should be in or below this directory. It should end with a slash. You should NOT include a filename.

Examples:
http://www.mysite.com/
http://www.mysite.com/products/
http://www.atelco.com/users/~joedoe/
https://intranet.bigcorp.com/finance/tax-ev

Name for this project:

This can be any unique name you want to give this project for your use. Usually this will be a part of your domain

Skip server check?

4. 點選「NEXT>」

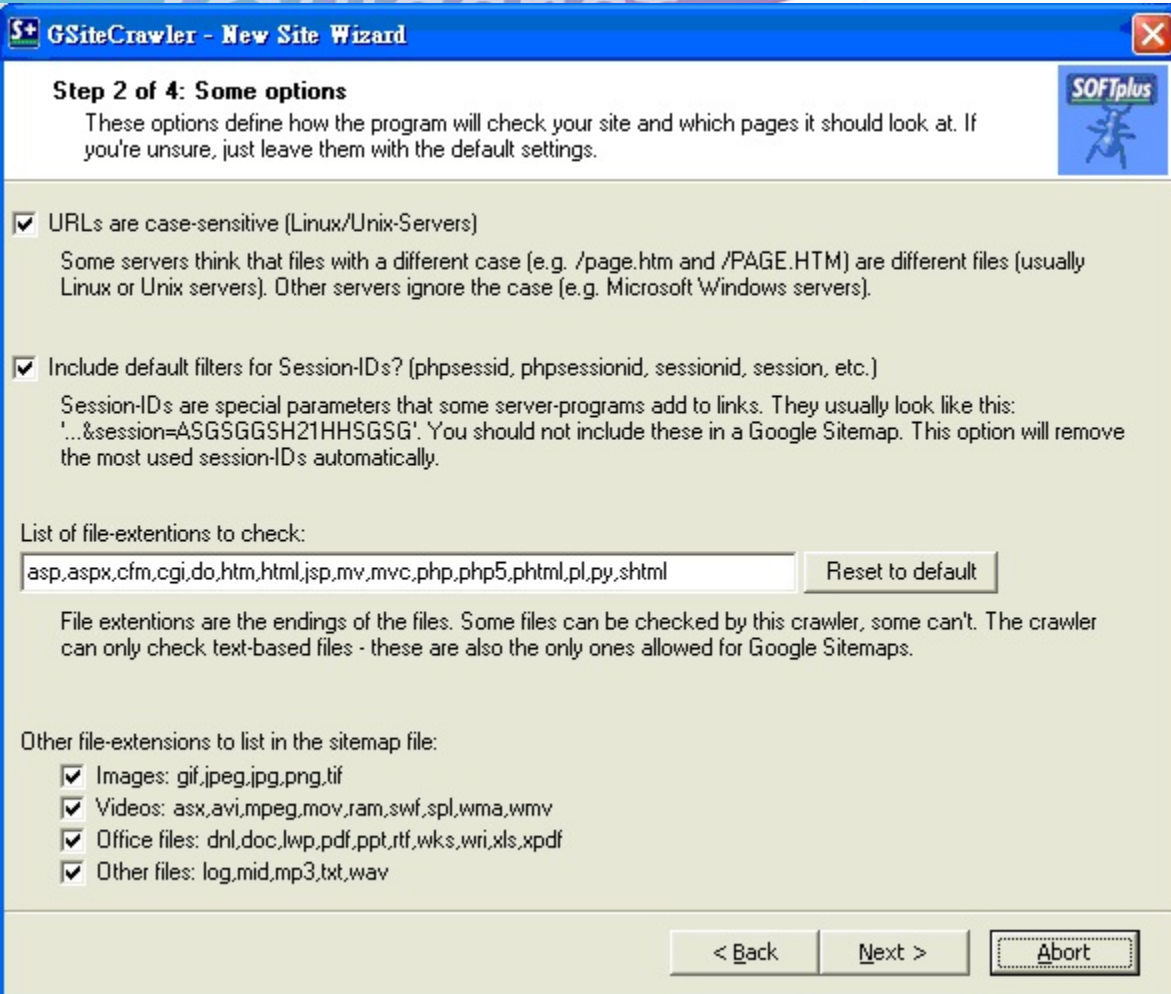
< Back **Next >** Abort

2. 填入所想檢視網站的主網址 (這裡以「台北商業技術學院－資訊管理系」網站為例：<http://imd.ntcb.edu.tw/>)。

3. 為該專案取名 (建議為操作者未來可辨識的名稱)，若不填寫，系統會以網址直接命名。

1. 點選「NEXT>」

加入專案



Step 2 of 4: Some options
These options define how the program will check your site. If you're unsure, just leave them with the default settings.

- URLs are case-sensitive (Linux/Unix-Servers)
Some servers think that files with a different case (Linux or Unix servers). Other servers ignore the case.
- Include default filters for Session-IDs? (phpsessid, etc.)
Session-IDs are special parameters that some servers use to track users. You should include the most used session-IDs automatically.

List of file-extensions to check:
asp,aspx,cfm,cgi,do,htm,html,jsp,mv,mvc,php,php5,...

File extensions are the endings of the files. Some servers can only check text-based files - these are also listed.

Other file-extensions to list in the sitemap file:

- Images: gif,jpeg,jpg,png,tif
- Videos: asx,avi,mpeg,mov,ram,swf,spl,wma,wmv
- Office files: dnl,doc,lwp,pdf,ppt,rtf,wks,wri,xls
- Other files: log,mid,mp3,txt,way

Step 3 of 4: Your FTP-Server
If you want to upload your sitemap file automatically, you can specify an FTP server.

- Upload sitemap file(s) via FTP?

Step 4 of 4: What can I do for you today?
These are some of the steps I can do for you today. I would suggest doing them all.

- Read ROBOTS.TXT
The file ROBOTS.TXT can be used to tell search engines where they shouldn't be looking. It's usually used to make sure that administrative pages aren't listed, etc. If you have a ROBOTS.TXT, it's important that you also exclude these areas in your sitemap file.
- Check for custom file-not-found error pages
Some servers have custom 'file-not-found' pages which look like normal pages. If a server has this, it can be hard for a search engine (or me) to tell if a page really exists or if it's missing. With this option I can check for that (but you'll need to tell me what part of the page is the error message).
- Import known pages from Google
Unless this is a really new site, Google will already know something about your pages. With this option, I'll check with Google first. Doing that can help the program get started faster.
- Scan your website now
If you're satisfied with these settings, I can start scanning your website right away.
- Wait for scan to complete and upload file
If you have a small site, you may wish to wait until I have everything scanned. After scanning your pages, I will upload the sitemap file (if you have specified the FTP server). And after that, I'll take you to the Google Sitemap website where you can either create an account or submit your sitemap.

1. 點選「Next >」

2. 點選「Next >」

3. 點選「FINISH」

檢查開始

情除下列所有詳細資料

「暫停/重新開始」

Pause Crawler Clear total queue Hide this window

Crawler	Status
Crawler 1	http://imd.ntcb.edu.tw/front/bin/cglist.phtml?Category=46
Crawler 2	http://imd.ntcb.edu.tw/front/bin/ptlist.phtml?Category=46
Crawler 3	http://imd.ntcb.edu.tw/front/bin/ptlist.phtml?Category=46
Crawler 4	http://imd.ntcb.edu.tw/front/bin/ptlist.phtml?Category=46
Crawler 5	http://imd.ntcb.edu.tw/front/bin/ptlist.phtml?Category=46
Crawler 6	http://imd.ntcb.edu.tw/front/bin/ptlist.phtml?Category=37

Crawler-Queue: Records waiting: 65, New/Min.: 115.0, Crawled/Min.: 12.0, Aborted/Min.: 0.0 (1 Min. avg)

URL	Date added	Errors
http://imd.ntcb.edu.tw/front/bin/rcglist.phtml?Rcg=4	2009/7/28 下午 04:32:	0
http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/257/214126272.xls	2009/7/28 下午 04:32:	0
http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/862199508.pdf	2009/7/28 下午 04:32:	0
http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/149181324.pdf	2009/7/28 下午 04:32:	0
http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/741704176.pdf	2009/7/28 下午 04:32:	0
http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/189876762.pdf	2009/7/28 下午 04:32:	0
http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/116172090.pdf	2009/7/28 下午 04:32:	0
http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/69810791.pdf	2009/7/28 下午 04:32:	0
http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/262/972nightclass.pdf	2009/7/28 下午 04:32:	0
http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/146047586.pdf	2009/7/28 下午 04:32:	0
http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/109825549.pdf	2009/7/28 下午 04:32:	0
http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/262/9756.pdf	2009/7/28 下午 04:32:	0

Aborted URLs: Right-Click to retry an aborted URL Show aborted list

URL	Error text	Linked from
-----	------------	-------------

隱藏該子視窗

(可再從主視窗功能表列  的叫出該視窗)

檢查中的連結
(項目)

檢查完成

The screenshot shows the Crawler-Watch application window. At the top, there are buttons for "Pause Crawler", "Clear total queue", and "Hide this window". Below these is a table with columns "Crawler" and "Status", listing Crawler 1 through Crawler 6. A dialog box in the center reads "The crawlers are now empty. (6)" with an "OK" button. Below the dialog, a progress bar shows "Crawler-Queue: Re..." and "1.3, Remaining: 00:00:00 (3 Min. avg)". The main area contains a table with columns "URL", "Date added", and "Errors". At the bottom, there is a section for "Aborted URLs:" with a "Show aborted list" button. A red box highlights this button, with a red arrow pointing to it from the text "以.txt文件檔檢視所有錯誤". Below the "Aborted URLs:" section is a table with columns "URL", "Error text", and "Linked from". A red box highlights this table, with a red arrow pointing to it from the text "該網站的錯誤項目".

The crawlers are now empty. (6)

The crawlers are now empty.
00:02:31

OK

Aborted URLs: Right-Click to retry an aborted URL

Show aborted list

URL	Error text	Linked from
http://imd.ntcb.edu.tw/front/bin/download.phtml?Part=970124&Nbr=7&Down=18	HTTP-Error 999 Page toc	http://imd.ntcb.edu.tw/front/bin/download.p
http://imd.ntcb.edu.tw/front/bin/ptdetail.phtml?Part=98051402&Rcg=4	HTTP-Error 999 Page toc	[GOOGLE]
http://imd.ntcb.edu.tw/front/mo/Mo8/style/style106/pictures/spacer.gif	HTTP-Error 404 Not Four	http://imd.ntcb.edu.tw/front/bin/viewcart.ph
http://imd.ntcb.edu.tw/www.google.com/calendar/embed?showTitle=0&showPri	HTTP-Error 404 Not Four	http://imd.ntcb.edu.tw/front/bin/home.phtml

該網站的錯誤項目

以.txt文件檔檢視所有錯誤

Show aborted list之txt檔

\$Aborted_20090728-1644 - 記事本

檔案(F) 編輯(E) 格式(O) 檢視(V) 說明(H)

Contents of aborted list in the current Crawler-Queue

2009/7/28 16:44 專案執行時間

Project: ntcb-imd 執行的專案名稱(第三頁中，使用者命名)

網站中有發生錯誤的頁面

URL: http://imd.ntcb.edu.tw/front/bin/download.phtml?Part=970124&Nbr=7&Down=1&Category=0
Error: HTTP-Error 999 Page too large, check settings
Linked from: http://imd.ntcb.edu.tw/front/bin/download.phtml?Part=970124&Nbr=7&Category=

URL: http://imd.ntcb.edu.tw/front/bin/ptdetail.phtml?Part=98051402&Rcg=4
Error: HTTP-Error 999 Page too large, check settings
Linked from: [GOOGLE] 網站錯誤類型

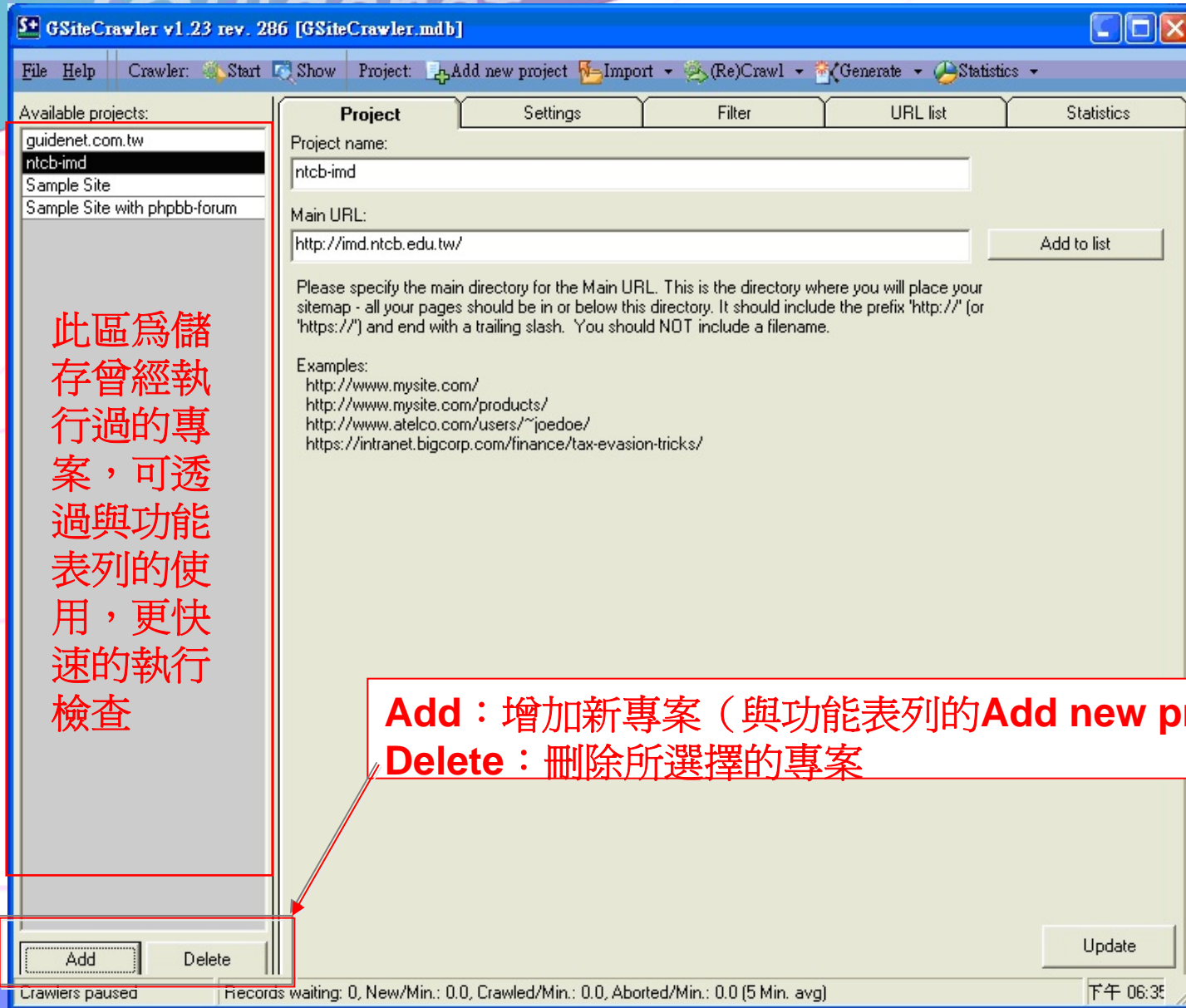
URL: http://imd.ntcb.edu.tw/front/mo/Mo8/style/style106/pictures/spacer.gif
Error: HTTP-Error 404 Not Found
Linked from: http://imd.ntcb.edu.tw/front/bin/viewcart.phtml 該錯誤網頁的來源網頁

URL: http://imd.ntcb.edu.tw/www.google.com/calendar/embed?showTitle=0&showPrint=0&showTz=08
Error: HTTP-Error 404 Not Found
Linked from: http://imd.ntcb.edu.tw/front/bin/home.phtml

Total aborted links in this project: 4 該網站(專案)共有多少錯誤

=====
Total aborted links: 4
=====
(End of file)

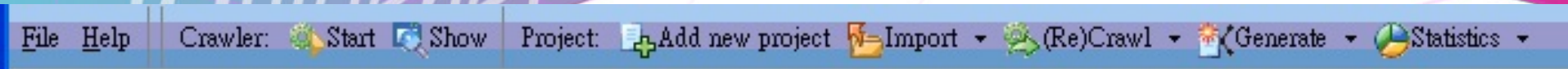
主視窗介紹



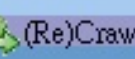
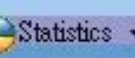


此區為儲存曾經執行過的專案，可透過與功能表列的使用，更快速的執行檢查

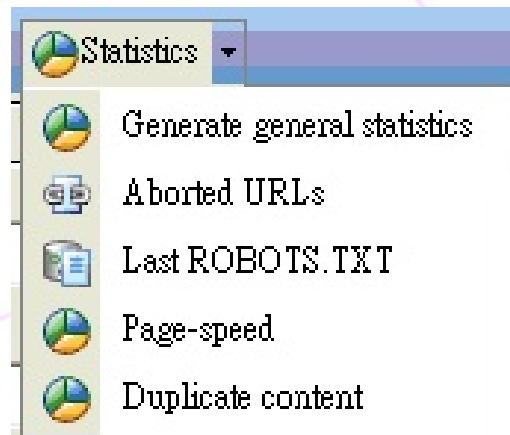
Add：增加新專案（與功能表列的Add new project相同功能）
Delete：刪除所選擇的專案

功能表列按鈕說明



-  Start :開始/暫停檢查
-  Show :檢視詳細檢查內容
-  (Re)Crawl :檢視詳細重新檢查
-  Statistics :檢視詳細的統計資料

該選項展開如:



以下將一一說明

統計資料



-Generate general statistics

Statistics for ntcb-imd Date: 2009/7/28 18:45

Main URL: <http://imd.ntcb.edu.tw/> (Case sensitive)

該網站的總網頁數

Number of URLs listed total: 163

Number of URLs listed to be included: 144

Number of URLs listed to be crawled: 78

總共檢查了多少網頁

Number of URLs still waiting in the crawler: 0 (may include some already listed)

Number of URLs aborted in the crawler: 4

還有多少
網頁仍等
待爬行

<http://imd.ntcb.edu.tw/front/bin/download.phtml?Part=970124&Nbr=7&Down=1&Category=0> linked from <http://imd.ntcb.edu.tw/front/bin/ptdetail.phtml?Part=98051402&Pcgs=4> linked from <http://imd.ntcb.edu.tw/front/bin/ptdetail.phtml?Part=98051402&Pcgs=4> linked from <http://imd.ntcb.edu.tw/front/bin/ptdetail.phtml?Part=98051402&Pcgs=4>
<http://imd.ntcb.edu.tw/front/mo/Mo8/style/106/pictures/spacer.gif> linked from <http://imd.ntcb.edu.tw/front/bin/ptdetail.phtml?Part=98051402&Pcgs=4>
<http://imd.ntcb.edu.tw/www.google.com/calendar/embed?showTitle=0&showPrint=0&showTz=0&mode=AGENDA&>

該網站的錯誤項目,以
linked from分開,左邊
為發生錯誤的網址,右
邊為該連結所在處(可
觀看**Aborted URLs**
統計資料較詳細)

Top 20 oldest included URLs:

<http://imd.ntcb.edu.tw/> (1483 days)
<http://imd.ntcb.edu.tw/front/pictures/spacer.gif> (1483 days)
<http://imd.ntcb.edu.tw/img/sysimg/empty.gif> (1483 days)
<http://imd.ntcb.edu.tw/front/pictures/rss.gif> (988 days)
<http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/255/ht11.gif> (797 days)
<http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/255/ht14.gif> (797 days)
<http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/109825549.pdf> (792 days)
<http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/116172090.pdf> (792 days)
<http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/128576352.pdf> (792 days)
<http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/132715470.pdf> (792 days)
<http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/149181324.pdf> (792 days)
<http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/150673171.pdf> (792 days)
<http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/163487019.pdf> (792 days)
<http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/189876762.pdf> (792 days)
<http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/196134285.pdf> (792 days)
<http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/204372918.pdf> (792 days)
<http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/208606125.pdf> (792 days)
<http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/567701862.pdf> (792 days)
<http://imd.ntcb.edu.tw/ezcatfiles/imd/img/img/269/69810791.pdf> (792 days)

該網站前 20 個較久沒更新的網頁

Ban URLs with these texts in them from being crawled:

Generate statistics

Open in Notepad

Generate statistics : 重新產生報表

Open in Notepad : 將該份報表以「.txt」檔觀看

統計資料

-Aborted URLs

Failed at 2009/7/28 17:17:

URL: <http://imd.ntcb.edu.tw/front/mo/Mo8/style/style106/pictures/spacer.gif>

Error: HTTP-Error 404 Not Found

Linked from: <http://imd.ntcb.edu.tw/front/bin/cglist.phtml?Category=2>

Failed at 2009/7/28 17:17:

URL: <http://imd.ntcb.edu.tw/www.google.com/calendar/embed?showTitle=0&showPrint=0&showTz=0&mod>

Error: HTTP-Error 404 Not Found

Linked from: <http://imd.ntcb.edu.tw/front/bin/home.phtml>

Failed at 2009/7/28 17:18:

URL: <http://imd.ntcb.edu.tw/front/bin/download.phtml?Part=970124&Nbr=7&Down=1&Category=0>

Error: HTTP-Error 999 Page too large, check settings

Linked from: <http://imd.ntcb.edu.tw/front/bin/download.phtml?Part=970124&Nbr=7&Category=>

Failed at 2009/7/28 17:18:

URL: <http://imd.ntcb.edu.tw/front/bin/ptdetail.phtml?Part=98051402&Rcg=4>

Error: HTTP-Error 999 Page too large, check settings

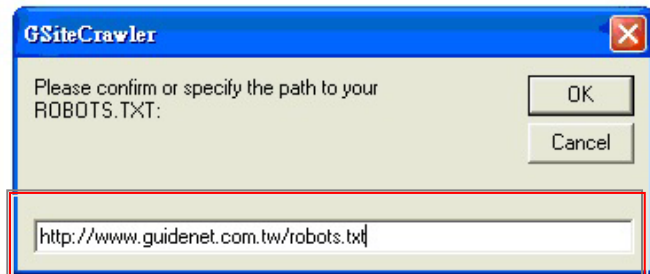
Linked from: <http://imd.ntcb.edu.tw/front/bin/rcglist.phtml?Rcg=4>

其詳細說明可見第八頁

-Last ROBOTS.TXT

檢視原始碼，因「台北商業技術學院-資訊管理系」的網址看不出什麼效果，在這裡改爲「知諸網科技股份有限公司」的網址。按下「**Generate statistics**」後：

No statistics, click below to generate them.



1.將「/robots.txt」刪除後按「ok」

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html>
<head>
<meta http-equiv="Content-Language" content="zh-tw">

<title>a??Noo?i$TaN?=?[3--??? q(Guidenet)-SEO/oo, oJa?P/oo, Au?A/-jM?TAo±MRa</title>

<meta name="description" content="a??Noo?i$T'£··N5-02£?~?PaA*EIG(1) oo, ^AEAu?AAU^YaA"
<meta name="keywords" content="SEO, -jM?TAoAu?A,oo, ^AE,oo, Au?A,oo, ?D?E,oo, o?@?

<style>
<!--
@import url("blog/Article/font.css");

body { font-family:"Verdana", "Arial", "Helvetica", "sans-serif"; font-size: 9pt}

p { font-family:"Verdana", "Arial", "Helvetica", "sans-serif"; font-size: 9pt}

a:link {
    text-decoration: none;
    color: #006699;
}

a:visited {
    text-decoration: none;
    color: #006699;
}

a:hover { color: #FF0033; text-decoration: underline}

td { font-family: "Verdana", "Arial", "Helvetica", "sans-serif"; font-size: 9pt}

.CompanyName {font-family: "Verdana", "Arial", "Helvetica", "sans-serif"; font-size: 23pt}
```

2.結果如此

Statistics for ntcg-imd Date: 2009/7/28 17:11

Main URL: <http://imd.ntcg.edu.tw/> (Case sensitive)

Top 30 largest pages:

<http://imd.ntcg.edu.tw/front/bin/ptdetail.phtml?Part=98052001&Rcg=4> (137.1 KB)
<http://imd.ntcg.edu.tw/front/bin/ptdetail.phtml?Part=98051401&Rcg=4> (133.7 KB)
<http://imd.ntcg.edu.tw/front/bin/ptdetail.phtml?Part=98061501> (84.8 KB)
<http://imd.ntcg.edu.tw/front/bin/ptdetail.phtml?Part=98061501&Rcg=1> (84.8 KB)
<http://imd.ntcg.edu.tw/front/bin/ptdetail.phtml?Part=98060901&Rcg=1> (67.7 KB)
<http://imd.ntcg.edu.tw/front/bin/ptdetail.phtml?Part=98060902&Rcg=1> (63.7 KB)
<http://imd.ntcg.edu.tw/front/bin/ptdetail.phtml?Part=980526> (63.6 KB)
<http://imd.ntcg.edu.tw/front/bin/ptlist.phtml?Category=57> (56.8 KB)
<http://imd.ntcg.edu.tw/front/bin/ptlist.phtml?Category=16> (55.6 KB)
<http://imd.ntcg.edu.tw/front/bin/ptdetail.phtml?Part=98052002&Rcg=4> (51.1 KB)
<http://imd.ntcg.edu.tw/front/bin/ptlist.phtml?Category=37> (49.4 KB)
<http://imd.ntcg.edu.tw/front/bin/ptlist.phtml?Category=64> (48.8 KB)
<http://imd.ntcg.edu.tw/front/bin/rcglist.phtml?Rcg=4> (48.8 KB)
<http://imd.ntcg.edu.tw/front/bin/rcglist.phtml?Rcg=1> (48.8 KB)
<http://imd.ntcg.edu.tw/front/bin/ptlist.phtml?Category=17> (47.5 KB)
<http://imd.ntcg.edu.tw/front/bin/home.phtml> (47.4 KB)
<http://imd.ntcg.edu.tw/front/bin/ptdetail.phtml?Part=980611&Rcg=1> (46.9 KB)
<http://imd.ntcg.edu.tw/front/bin/ptdetail.phtml?Part=980611&Rcg=4> (46.9 KB)
<http://imd.ntcg.edu.tw/front/bin/ptdetail.phtml?Part=980514&Rcg=4> (44.1 KB)
<http://imd.ntcg.edu.tw/front/bin/ptdetail.phtml?Part=98061101&Rcg=1> (43.7 KB)
<http://imd.ntcg.edu.tw/front/bin/ptdetail.phtml?Part=98061101&Rcg=4> (43.7 KB)
<http://imd.ntcg.edu.tw/front/bin/ptlist.phtml?Category=20> (43.5 KB)
<http://imd.ntcg.edu.tw/front/bin/ptlist.phtml?Category=30> (43.0 KB)
<http://imd.ntcg.edu.tw/front/bin/ptdetail.phtml?Part=98061102&Rcg=1> (42.9 KB)
<http://imd.ntcg.edu.tw/front/bin/ptlist.phtml?Category=34> (42.8 KB)
<http://imd.ntcg.edu.tw/front/bin/ptlist.phtml?Category=24> (42.5 KB)
<http://imd.ntcg.edu.tw/front/bin/ptlist.phtml?Category=23> (42.2 KB)
<http://imd.ntcg.edu.tw/front/bin/ptdetail.phtml?Part=980608&Rcg=1> (41.7 KB)
<http://imd.ntcg.edu.tw/front/bin/ptdetail.phtml?Part=980608&Rcg=4> (41.7 KB)

Top 30 slowest pages to download:

<1>Top 30 largest pages:

該網站中網頁容量前 30 大的網頁
(以容量大小表示)

<2>Top 30 slowest pages to download:

該網站中前 30 下載最慢的網頁
(以總花費時間表示)

<3>Top 30 slowest pages to download (Size/Time):

該網站中前 30 下載最慢的網頁
(以下載速度(Size/Time)表示)

<4>Top 30 slowest pages to crawl (many links):

該網站中前 30 爬行最慢的網頁
(表示該網頁連結較多)

-Duplicate content

檢視相同（相似）內容的網頁（表示其中有些網頁是以舊網頁直接複製再略為修改而成）

Statistics for ntcg-imd Date: 2009/7/28 17:12

Main URL: <http://imd.ntcb.edu.tw/> (Case sensitive)

URL's with duplicate content (identical pages):

<http://imd.ntcb.edu.tw/front/bin/ptdetail.phtml?Part=98052201> (e*s?0*_0~\$T3N?C*, eToT2z*~t--li?x?~ljak
<http://imd.ntcb.edu.tw/front/bin/ptdetail.phtml?Part=98052201&Rcg=4> (e*s?0*_0~\$T3N?C*, eToT2z*~t--li;

<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=980505> (ptclick)
<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=980514> (ptclick)
<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=98051401> (ptclick)
<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=98052001> (ptclick)
<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=98052002> (ptclick)
<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=980521> (ptclick)
<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=98052201> (ptclick)
<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=980526> (ptclick)
<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=980604> (ptclick)
<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=980608> (ptclick)
<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=980609> (ptclick)
<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=98060901> (ptclick)
<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=98060902> (ptclick)
<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=980611> (ptclick)
<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=98061101> (ptclick)
<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=98061102> (ptclick)
<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=98061501> (ptclick)
<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=980617> (ptclick)
<http://imd.ntcb.edu.tw/front/bin/ptclick.phtml?Part=98063001> (ptclick)

End of data.

「<http://imd.ntcb.edu.tw/front/bin/ptdetail.phtml?Part=98052201>」
與「<http://imd.ntcb.edu.tw/front/bin/ptdetail.phtml?Part=98052201&Rcg=4>」
就是內容相同（相似）的網頁