



**Inspyder Web2Disk
User's Reference Manual**

Copyright © 2008 Inspyder Software Inc.

Inspyder Web2Disk (W2D) User's Reference Manual

Copyright © 2008 Inspyder Software Inc.

All rights reserved. No parts of this work may be reproduced in any form or by any means - graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems - without the written permission of Inspyder Software Inc.

Products that are referred to in this document may be either trademarks and/or registered trademarks of the respective owners. The publisher and the author make no claim to these trademarks.

While every precaution has been taken in the preparation of this document, the publisher and the author assume no responsibility for errors or omissions, or for damages resulting from the use of information contained in this document or from the use of programs and source code that may accompany it. In no event shall the publisher and the author be liable for any loss of profit or any other commercial damage caused or alleged to have been caused directly or indirectly by this document.

Printed May 2008 in Canada.

Table of Contents

Part 1 Introduction

Part 2 Quick Start Guide

Part 3 Usage

3.1	Button Bar & Menu	3
3.2	Project Settings	4
3.2.1	Advanced Project Settings	6
3.2.2	Import Robots.txt	8
3.3	Crawl Results	9
3.4	Settings	10
3.4.1	Scheduler	10
3.4.2	Email Settings	12
3.4.3	Master Project Settings	13

Part 4 About Inspyder Software

Part 5 Product License Agreement

Index

1 Introduction

Web2Disk from [Inspyder Software Inc.](#) is a Windows based utility that lets you download an entire site, including images, style sheets and other embedded content, to your PC for offline browsing. Web2Disk automatically updates the content it downloads to allow offline browsing. Web2Disk can download dynamic websites by automatically adjusting filenames and internal links. Even if the site you want to download is complex, Web2Disk can handle it.

Web2Disk is a web crawler so it saves the pages that your browser would get when browsing normally. Web2Disk is compatible with Apache, IIS and other web server software. It can download websites created with PHP, ASP, JSP or any other technology. Just enter a URL and let Web2Disk do the rest.

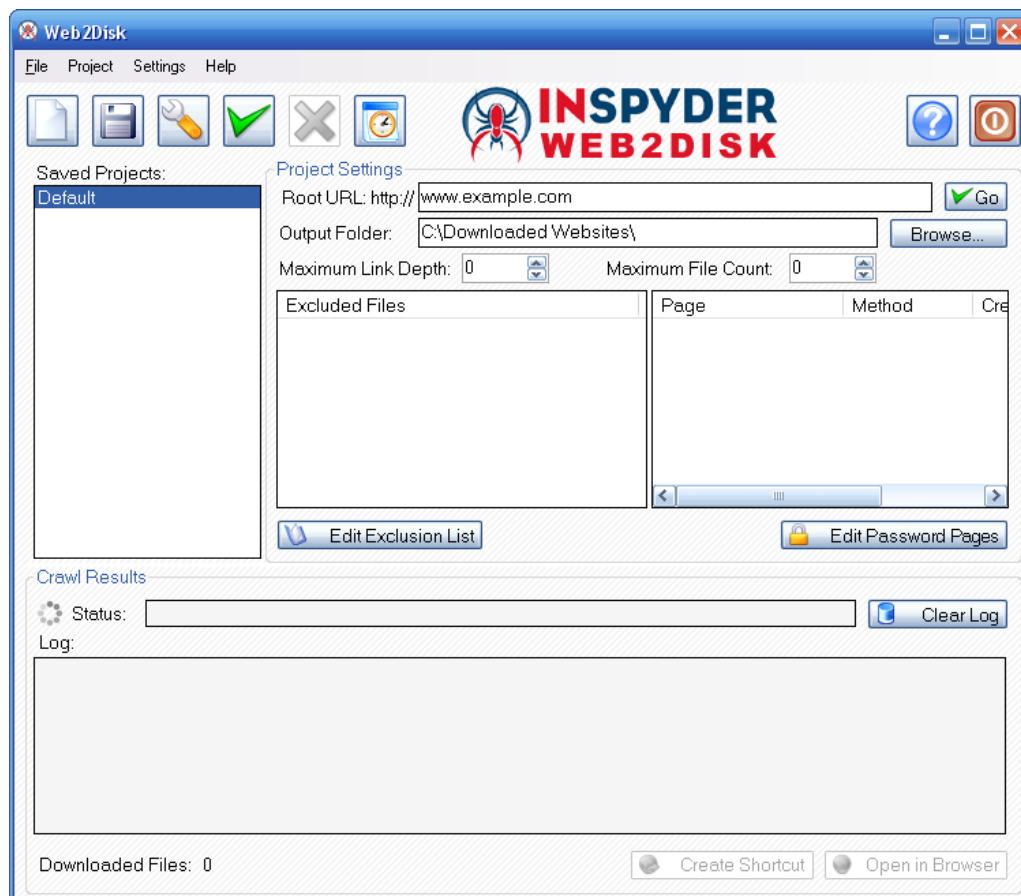
Features and Benefits

- Easy to Use — Just enter a website URL and go!
- Offline Browsing — Web2Disk fixes downloaded content for easy offline browsing.
- Scheduled Website Downloads — Take a website with you, where no Internet connection is available!
- Monitor a Website for Updates — Configure Web2Disk to email you when a site is changed.
- Download Dynamic Pages — Download database driven websites with ease. Web2Disk converts dynamic content to static content for offline browsing.
- Powerful Filtering — Save bandwidth by excluding the files, pages and folders you don't need.

Web2Disk requires Windows 2000/XP/2003. Please refer to the [web site](#) for current licensing and pricing information.

2 Quick Start Guide

- Step 1** Enter the URL of the website you wish to save in the 'Root URL' field.
Step 2 Enter the folder you wish to save the website to in the 'Output Folder' field.
Step 2 Click the "**Go**" button.
Step 3 When crawling is finished, click 'Open in Browser' or 'Create Shortcut' to see the saved website!



3 Usage

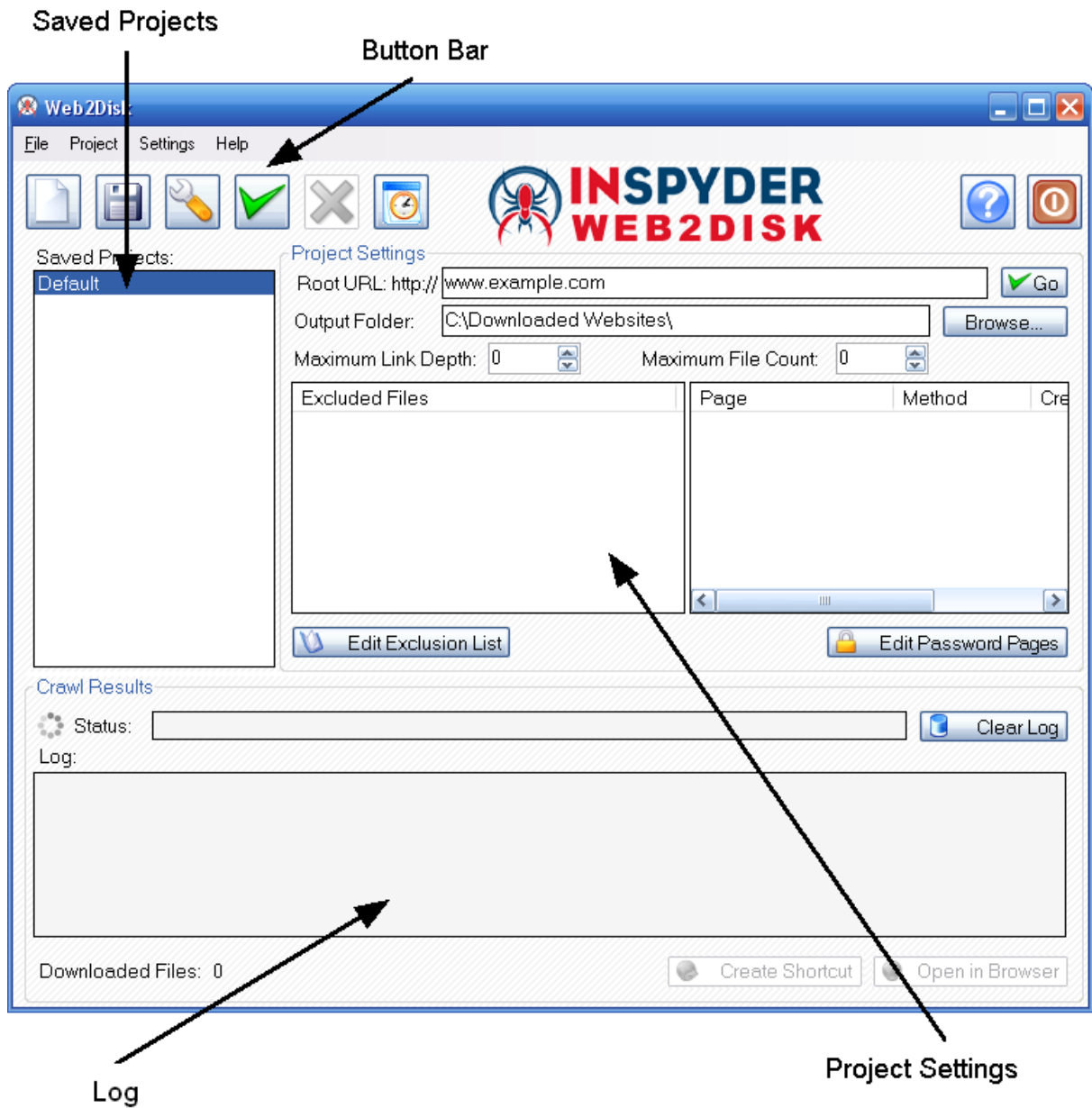
The Web2Disk user interface area is broken down into four main areas:

Button Bar The button bar contains shortcuts to the most used features of Web2Disk.

Saved Projects This is a list of the currently saved projects. When you create a new project, it will appear in this list. Projects can be deleted or renamed by right clicking the project name in the project list.

Project Settings This is where the settings for the currently select project will appear.

Crawl Log This is where the output from Web2Disk will be presented.



The following sections describe the individual fields and controls of Inspyder Web2Disk's user interface.

3.1 Button Bar & Menu

This section describes the functions of the button bar controls. Some of this functionality is available by right-clicking on controls in other parts of the user interface.



The **New Project** button is used to establish a new web site that you wish to analyse. You will be prompted for a shortcut name, which will appear in the Projects list, and the root URL of that web site.



The **Save Project** button is used to save the current project settings. The main menu contains a "**Save As**" command if you want to replicate the current project. As well, you can right-click on any project name in the Saved Projects list to delete or rename the project.



This option allows you to quickly access the Advanced Project Settings. For more information please refer to the [Advanced Project Settings](#) section of this document.



Select a project from your list of previously defined projects and then click on this button to start the download process. The crawling can be interrupted at any time by clicking on the **Stop** button.



Use this button to stop the crawling process.



This button provides a shortcut to [Schedule](#) your Web2Disk operation.



Provide access to the help file and About screen. The About screen contains version and your registration information.



Closes the program.


3.2 Project Settings

This section describes the options, fields and features of the Project Settings group.

Root URL

Specify the address of the homepage for the website to be downloaded. You can enter an IP address or the URL of the web page. The URL does not need to begin with the "http://" prefix. (If your site requires SSL, include the "https://" prefix.)

Go Button

Use this control to begin the download from the Root URL. This button has the same functionality as the  control on the button bar. It's included in this location for convenience.

Output Folder

Specify the location where the newly downloaded website is to be stored.

Browse Button

Use this control as a convenient way to locate the Output Folder.

Maximum Link Depth

Limits how "deep" into the site Web2Disk should crawl. A value of 1 represents every file linked off the first page ("1 click"). A value of 2 would represent any files accessible by following 2 links into the site. Set this value to zero for no limit.

Maximum File Count

Limits the total number of files that Web2Disk should download from the current

site. Once this number of files is downloaded, Web2Disk will stop crawling. Set this value to zero for no limit.

Excluded Files

The Excluded Files listbox contains URLs that should not be part of the analysis. This is important if you have pages or scripts that generate a large amount of unimportant content (such as a message board or article pages that are formatted for printing). You might not want these included in your download.

When a directory (like "/blog") is specified as an Excluded Path, then all of its sub-directories will be excluded (such as: `www.mysite.com/blog/index.html` and `www.mysite.com/blog/archives/january2005.html`).

Wildcard matching is also supported through the use of the '*' (asterisk) character. The '*' will replace zero or more characters as shown in the examples below. The '*' can be used multiple times in a string.

<i>inventory/warehouse*index</i>	will exclude the following: <i>inventory/warehouse1index</i> <i>inventory/warehouse125index</i> <i>inventory/warehouseindex</i>
<i>userreports/*.txt</i>	will exclude all files with an extension of ".txt"

Click the **Edit Exclusion List** button to add URLs to the exclusion list. Internal URLs should be relative to the root URL. URLs beginning with '/' are taken from the domain root (this is not recommended). URLs **not** beginning with '/' are taken from the start of the crawl. For example, if the root URL is "www.mysite.com/somepage/":

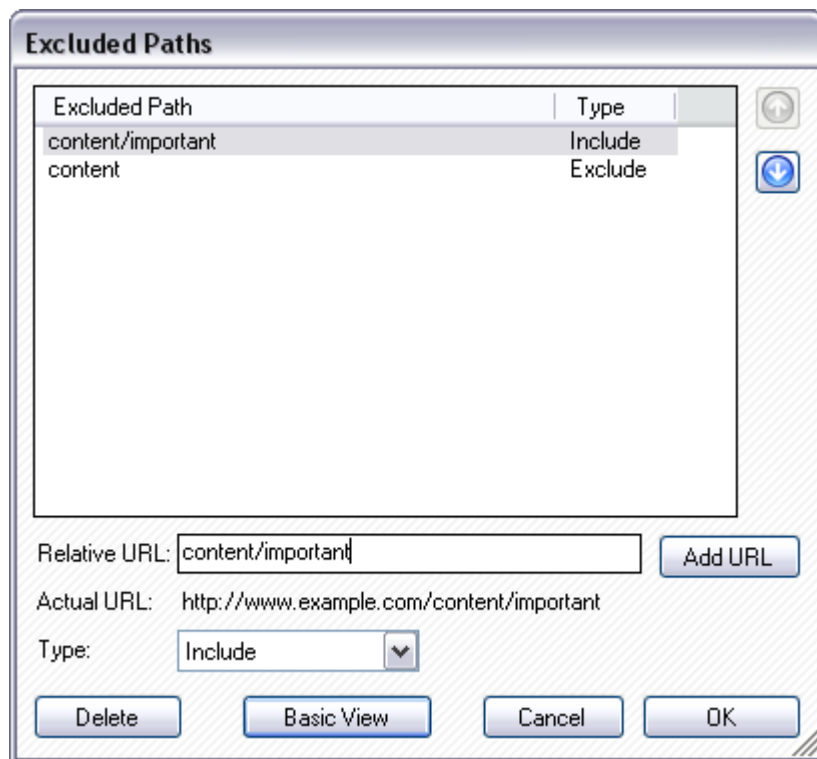
'apage.html' becomes *'www.mysite.com/somepage/apage.html'*
'/apage.html' becomes *'www.mysite.com/apage.html'*

URLs beginning with 'http://' are treated as full URLs, and can be used to exclude links to external sites. For example:

'http://www.anothersite.com' **remains** *'http://www.anothersite.com'*

A page can also be added to the exclusion list by right clicking the URL in the Results tab.

By clicking the "Advanced View" button on the exclusion dialog you are able to mark a page or section as specifically **included**. The list is matched from top to bottom. If you want to exclude all pages under "www.example.com/content/" except "www.example.com/content/important" then you would create an entry that includes "/content/important" above the rule that excludes "/content". You can move the rules up and down with the arrows on the right.



Password Pages

Password protected pages can be configured by clicking the **Edit Password Pages** button. A dialog box will appear that will let you enter the relative URL of the password protected page, as well as the authentication method and the credentials. If the HTTP method is selected then a username, password and domain can be entered into the appropriate fields. Note that the domain field is optional.

If the POST method is selected then the credentials must be entered as form data. (For example, the following form data can be posted to a page: *username=user&password=pass*. Where 'username' and 'password' are the form items, and 'user' and 'pass' are the values.) To automate this, click the 'Wizard' button and use the built-in browser to surf to the site you wish to log into. When you log in, Web2Disk will automatically capture the form data submitted to the site.

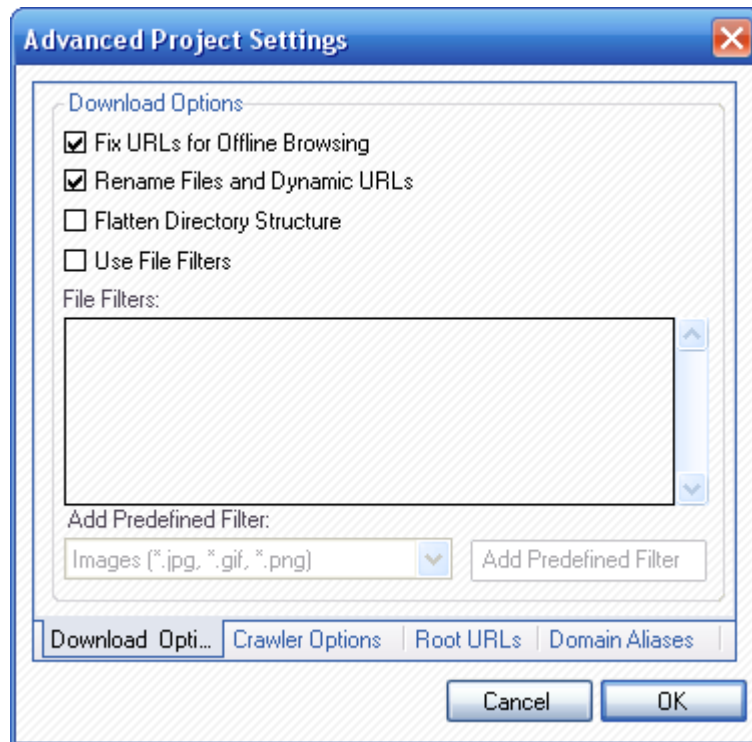
Refer to the Excluded Pages description above for examples of proper URL formation.

For advanced users it may be of value to note that Web2Disk Projects and settings are stored in folder:

C:\Documents and Settings\username\Application Data\Inspyder Web2Disk\

3.2.1 Advanced Project Settings

The Advanced Project Settings provide you with a way to further refine how Web2Disk crawls a website. These settings are project specific.



Download Options

Fix URLs for Offline Browsing

If this option is checked Web2Disk will go through the downloaded content and re-write any internal links so that they point to the downloaded files on your hard drive. If this options is unchecked, the downloaded content will be left "as-is", offline browsing may or may-not work. The default value is checked.

Rename Files and Dynamic URLs

If this option is checked Web2Disk will automatically rename file extensions so that Windows makes the correct file-type association when opening the offline files. Dynamic URLs (with parameters, such at <http://www.example.com/products.aspx?productID=xyz&view=normal>) are also renamed to over come any file name limitations of Windows.

If this option is unchecked, the files will be saved with the same names used by the server, but dynamic URLs will no longer be downloaded. Additionally, Windows may fail to open some file types correctly. The default option is checked.

Flatten Directory Structure

If this option is checked, all the downloaded files will be stored in the same folder. Files with duplicate names, such as:

- <http://www.example.com/index.php>
- <http://www.example.com/support/index.php>

will be over-written by the last file with that name that is downloaded. This option is useful if you wish to strip the content files from a site (such as images, CSS, etc.) but are not interested in the HTML files.

If this option is selected, it is not possible to use the 'Fix URLs for Offline Browsing' feature.

Use File Filters

File Filters are used to restrict what Web2Disk actually downloads and saves. (The Exclusion List is used to filter how Web2Disk crawls a website.) You can use File Filters if you only want to download a certain type of file from a website. For example, to download only PDF documents, you would enter "*.pdf" on a line in the File Filters.

Multiple filters can be placed one per line, or on a single line separated by commas. The * character is used to match one or more letters. For example, "*.jpg" would match any files that end with ".jpg".

Crawler Options

Crawler Pacing

Crawler Pacing sets the delay between each file that is downloaded from a website. If Web2Disk is crawling a site too fast, increase this value to slow it down.

Crawler Timeout

The Crawler Timeout is the number of seconds that Web2Disk should wait for a response from the remote server. If the server does not respond within the set number of seconds, the file is skipped.

User Agent

The User Agent field sets the "User Agent" value that Web2Disk will use to identify itself to remote servers. All web-browsers and crawlers have this ability. By default, Web2Disk uses the Inspyder user agent string. If a particular site rejects Web2Disk, it's possible to change the User Agent string so that Web2Disk masquerades as Internet Explorer or FireFox.

Crawl JavaScript

If this option is checked, Web2Disk will attempt to follow certain types of links made in JavaScript.

Additional Root URLs

Additional Root URLs are used to 'seed' the crawler if your site is not fully cross-linked. If you have doorway pages, or a disconnected section of your site (to which no links point), you can enter these pages as Additional Root URLs. The crawler will process these pages while crawling your site.

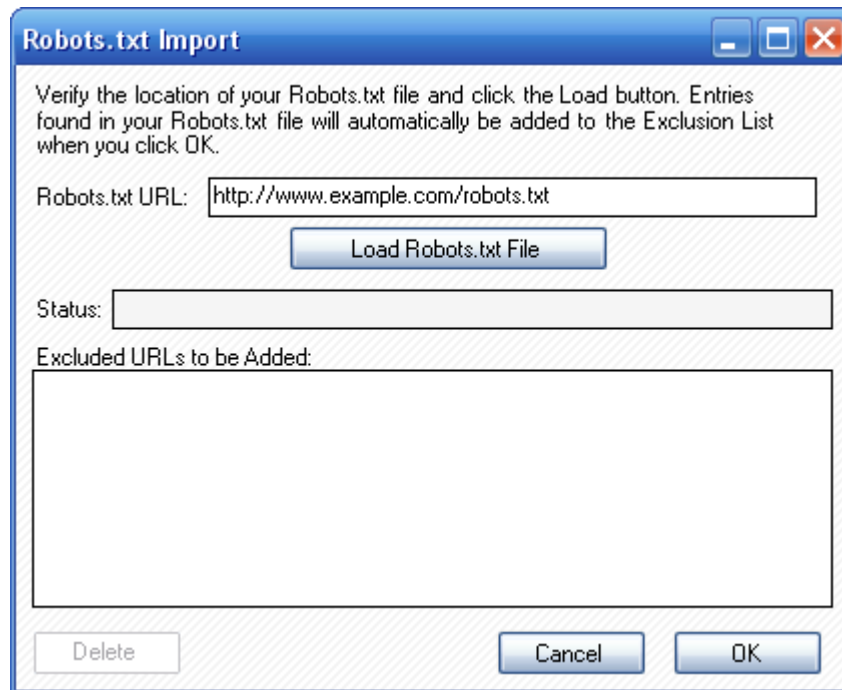
To add a new Root URL, simply enter it in the URL field and click 'Add'. To remove a URL from the list, select it and press the delete key ('Del') on your keyboard.

Domain Aliases

Domain aliases are used to tell the crawler that your site has more than one domain name. This feature is useful if your site is referenced internally by more than one domain. For example, many sites use "www.example.com" and "example.com" interchangeably. If the root URL has the "www" prefix, links using the other domain name will be treated as external to the site (and ignored) unless it is specified as a domain alias.

3.2.2 Import Robots.txt

If your website uses a "robots.txt" file which contains URLs you wish to exclude Web2Disk from crawling, you can import them with the Robots.txt Import tool. To access this window, select "Project | Import Robots.txt..." from the main menu bar.

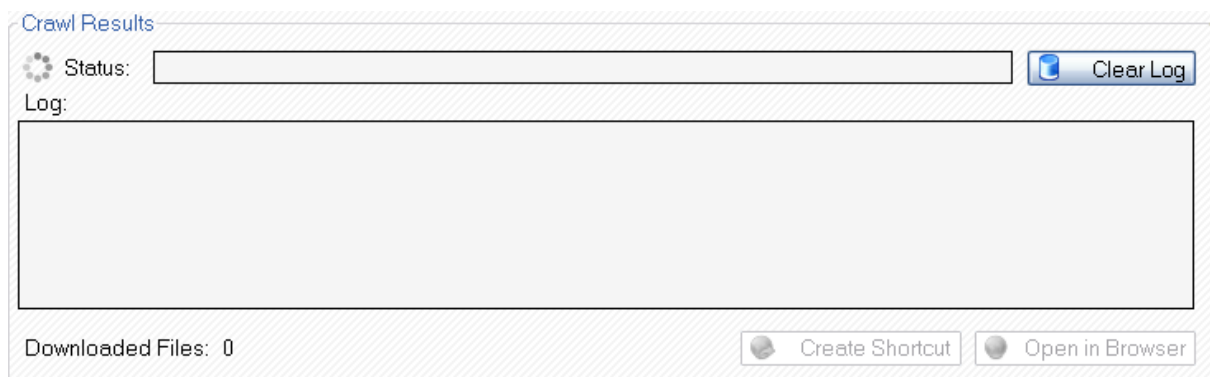


The "Robots.txt URL" field should contain the location of your website's robots.txt file. Web2Disk will try and guess this URL based on the Root URL for the current Project. If the URL that Web2Disk guesses is not correct, you can correct it in place. To load the file, click "Load Robots.txt File". Web2Disk will retrieve the file from your server and include any rules that match the wildcard User Agent ("User-agent: *"). Other rules defined explicitly for certain robots/crawlers will be ignored.

Any rules that are found in the Robots.txt file will appear in the "Excluded URLs to be Added" list. You can remove any rules you don't want to import by selecting them and clicking the 'Delete' button.

3.3 Crawl Results

This section describes the options, fields and features of the Crawl Results group.



Status

The status line shows the current file being downloaded by Web2Disk. If a file takes more than 1 second to download, Web2Disk will also show the current download progress and speed in this line.

Clear Log

The Log provides a history of all the events and errors that occur while crawling a site. If you are experiencing difficulty with Web2Disk, the log can be helpful in trouble shooting your problem. Clicking the 'Clear Log' button will remove all the entries.

Create Shortcut

This button provides a quick way to create a Windows shortcut to the downloaded website so it can be easily browsed later. This button is disabled until your crawl completes successfully.

Open in Browser

This button provides a shortcut to opening the newly downloaded website in your default browser. This button is disabled until your crawl completes successfully.

3.4 Settings

3.4.1 Scheduler

Web2Disk has the ability to run in an unattended mode and automatically download a website and check that site for changes since the last download. To configure Web2Disk for scheduled operation simply click the '**Clock**' button from the main button bar. The dialog below contains the choices from the as well as those necessary for timing the execution of Web2Disk. Once the Schedule Options are filled in and the **Add Task** button is clicked the task will be scheduled using the Windows Task Scheduler.

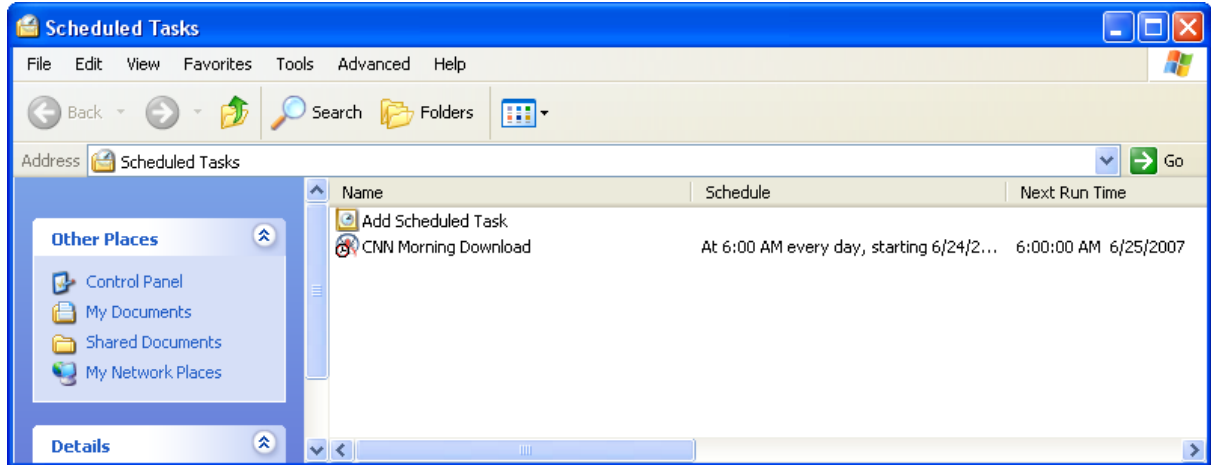
Send Email Notification

If this option is checked, Web2Disk will automatically send an email to email address specified when the website has been downloaded.

Only Send Notification if Website Changes

If this option is checked, the email notification will only be sent if changes are detected on the downloaded site. The first download will always result in an email being sent since all pages will appear new to Web2Disk. This option is extremely useful if you wish to be notified when a particular site or page has been updated by it's owner.

Use the **View Schedule** button to view all the tasks submitted to the Windows Task Scheduler as well as **to modify the scheduling parameters**. If you double-click the scheduled entry in the Windows Scheduled Tasks dialog a new dialog will be displayed which contains all of the parameters for that task.



The actual command used to invoke Web2Disk is found in the 'Run:' field. You can modify the parameters used by Windows when it starts Web2Disk by **carefully** editing the contents of the Run field. The following are the command line arguments:

-project=	The project name to be downloaded (i.e. -project=default)
-rooturl=	(Optional) Can be used to override the project's Root URL
-filelimit=	(Optional) Overrides the project's File Limit setting
-depthlimit=	(Optional) Overrides the project's Depth Limit setting
-email=	(Optional) The email address to use for notification
-checkchanges	(Optional, requires -email) Indicates that an email notification should only be sent if changes are detected

To satisfy the Task Scheduler's parsing algorithm please ensure that each argument is placed within double-quotes (i.e. "-rooturl=www.example.com"). This is usually only required if the argument contains spaces.

Note to Windows XP and 2003 users: The scheduled tasks will not run on an account with a blank password. For a workaround to this please see the Web2Disk FAQ.

Scheduled Execution Return Codes:

These are the codes returned to the Windows Task Scheduler after the Web2Disk task has been executed. You can view them in the Scheduler window to help troubleshoot any problems.

0	Success (no errors)
1	Not registered Solution: Run the program interactively and enter your registration information when prompted.
2	Configuration Error Solution: A setting in the software is incorrect (such as an invalid save path). Try running the software in interactive mode, and check for errors.
3	Crawler Error Solution: The crawler detected some type of unhandled error during crawling. Run this project in interactive mode and check for errors (corrupt HTML, etc.).

3.4.2 Email Settings

Web2Disk has provisions for customizing the emails that can be sent when scheduled scanning has completed. The customization screen is accessed by clicking on the "**Tools | Customize Email**" menu item.

Enter the **SMTP Server** (mail server) that is used for your outgoing email, and fill out the **Sender** field with the email address to be used as the sender. These settings will be used when automatically exporting the results of a [scheduled analysis](#). The recipient's address is setup when setting Web2Disk to run on a scheduled basis.

Simply modify the **Subject** line or the email **Body** as desired and click the **OK** button to save your changes. Special tags can be used which are replaced with specific values by Web2Disk before the email is sent.

The following tags can be used within the email Body and/or Subject line. (Tags are not case sensitive.)

Tag	Location	Description
#project#	Subject and Body	The project name
#rooturl#	Subject and Body	The Root URL of the project
#offlineurl#	Body	The location where the site was saved on your PC
#changes#	Body	A listing of the files that changed since the last crawl

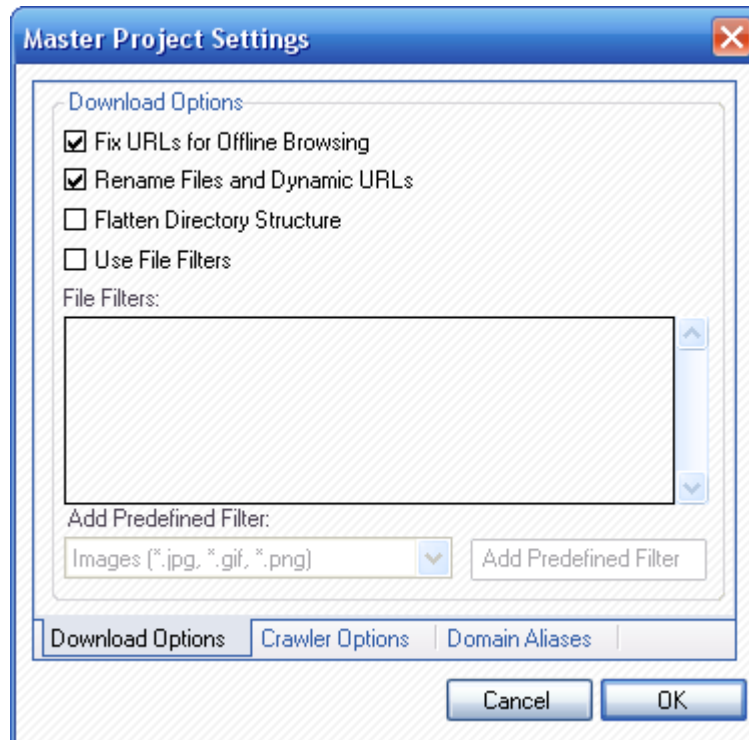
The sample screen illustrates the usage of the tags in customizing the email message **Subject** and **Body**.

The screenshot shows the "Email Settings" dialog box. It includes the following fields and controls:

- SMTP Server:** Text input field.
- Use SSL:** Unchecked checkbox.
- Username:** Text input field with "(Optional)" label.
- Password:** Text input field with "(Optional)" label.
- Sender Address:** Text input field.
- Subject:** Text input field containing "#Project# Website Download Report".
- Body:** Text area containing "#Project# has been successfully downloaded by Web2Disk. To view the downloaded copy, open the following URL if your browser: #offlineurl#" and "The following changes were detected: #CHANGES#".
- Buttons:** "Default", "Test", "Cancel", and "OK".

3.4.3 Master Project Settings

In addition to the regular user defined Projects, Inspyder Web2Disk contains a single Master Project. The Master Project contains the default values for any newly created projects. Options that are unique to every Project (such as Root URL) cannot be defined in the Master Project. To edit the Master Project click 'Tools | Edit Master Project...'. The settings in the Master Project Settings window are identical to those of a regular project.



For details on what each specific options means, please review the settings in the ['Advanced Project Settings'](#) section of the manual.

4 About Inspyder Software

Founded in 2004, [Inspyder Software](#) is a leading provider of web crawling technologies for content analysis, filtering and web access management. These technologies have formed the foundation of our software products and made 'Inspyder' a recognizable brand in the software industry.

As the World Wide Web continues to grow, organizations are looking for ways to improve their online presence, increase the quality of their sites, and maximize overall user experience. At Inspyder Software, our focus is on the development of technology that enables organizations to do just that.

For more information regarding future products, custom development or consulting services, please contact us at sales@inspyder.com.

5 Product License Agreement

This legal document is an agreement between you, as licensee and Inspyder Software Inc. ("Inspyder"), as licensor. You should carefully read the following terms and conditions before using this product. Using this product indicates your acceptance of these terms and

conditions.

DEFINITIONS: "Product" means (a) all of the contents of the files, disk(s), CD-ROM(s) or other media with which this Agreement is provided, including but not limited to Inspyder or third party computer information or software; written documentation or documentation files; and (b) upgrades, modified versions, updates, and additions. "Use" or "Using" means to access, install, download, copy or otherwise benefit from using the functionality of the Product in accordance with the documentation. "You" and "Your" means the purchaser of the Product.

GRANT OF LICENSE: The Licensor grants to You a non-transferable non-exclusive license to use the product on a single computer. You may physically transfer the product to more than one computer provided the product is used only on a single computer. You may not modify, adapt, translate, reverse engineer, decompile, disassemble or create derivative work based on: (a) the product; (b) written material associated with the product; (c) the concepts or technology utilized in this product.

COPY RESTRICTIONS: You may not copy the product including any product that has been modified, merged or included with other products except as specified in this Agreement, nor may You copy any written materials associated with the product. You shall be held legally responsible for any copyright infringement that is caused or encouraged by Your failure to abide by the terms of this license.

TRANSFER RESTRICTIONS: The product is licensed only for You, and You may not transfer the product or the license to use the product without Licensor's prior written consent. Any authorized transferee of the product shall be bound by the terms of this Agreement. In no event may You transfer, assign, rent, lease, sell or otherwise dispose of the product on a temporary or permanent basis except as expressly provided for herein. The product cannot be resold.

USAGE RESTRICTIONS: The product is licensed only for Your personal use. You cannot use the product to provide a service to others. To be clear, you can use the product as part of your web development activities, including web development for others. However, you cannot use Web2Disk to provide an exclusive service such as using Web2Disk as part of a web checking or analysis service.

LIMITED WARRANTY: Except as expressly set forth herein, the product is provided 'AS IS' without warranty of any kind, either express or implied, including, but not limited to the implied warranties of merchantability and fitness for particular purpose. Inspyder Software Inc. does not warrant that the function of this product will be error free. However, Inspyder Software Inc. does warrant the media on which the software is furnished to be free from defects in material and workmanship under normal use for a period of 30 days from the date of delivery to You.

LIMITATIONS OF REMEDIES: The Licensor's entire liability and Your exclusive remedy shall be: (a) the replacement of any media not meeting Licensor's 'Limited Warranty' and which is returned to Licensor or an authorized representative of Inspyder Software Inc. with a copy of Your receipt; or (b) if Inspyder Software Inc. is unable to deliver replacement media which is free of defects in materials or workmanship, You may terminate this Agreement by returning the product and Your money will be refunded. In no event will Licensor nor anyone involved in the creation, production or distribution of the product be liable to You for any direct, indirect, consequential or incidental damages including any lost profits, lost savings, lost business revenue or other commercial or economic loss arising out of the use or inability to use the product even if Licensor or any authorized representative of Licensor has been advised of the possibility of such damages or for any claim by any other party.

JURISDICTION: The laws of the Province of Ontario, Canada shall govern this Agreement.

All rights are reserved; Copyright © 2008 Inspyder Software Inc.

Index

- A -

Advanced Project Options 1

- B -

Button Bar 2, 3

Buttons

Exit Program 3

Export 3

Help 3

New Project 3

Save Project 3

Start Crawl 3

Stop Crawl 3

- D -

Default Values 6

- E -

Email Settings

Body 12

Sender 12

SMTP Server 12

Subject 12

Excluded Files 4

Exit Program Button 3

Export Button 3

- F -

Features and Benefits 1

FTP Options 6

- H -

Help Button 3

- I -

Inspyder Software Inc. 13

Internet Usage Management 13

Introduction 1

- M -

Multiple Root URLs 6

- N -

New Project Button 3

- O -

Options 9

Overview 2

- P -

Product License

Copy Restrictions 13

Definitions 13

Grant of License 13

Jurisdiction 13

Limitations of Remedies 13

Limited Warranty 13

Transfer Restrictions 13

Project Settings 2

Exclusion List 4

Root URL 4

- R -

Root URL 4, 6

- S -

Save As 3

Save Project Button 3

Saved Projects 2

Scan Results 2

Log 9

Scan Results	2
Sitemap Info	9
Scheduled Execution Return Codes	10
Site Sections	4
SMTP Server	12
Start Analysis Button	3
Stop Analysis Button	3



Inspyder Software Inc.
698 Holt Drive,
Burlington, Ontario,
Canada, L7T 3N5

Tel: 888.732.6134

Web: <http://www.inspyder.com>
Email: support@inspyder.com