Inspyder Power Search
User's Reference Manual

Inspyder Power Search (Power Search)
User's Reference Manual

Copyright © 2014 Inspyder Software Inc.

Printed October 2014 in Canada.

Table of Contents

1   Introduction

Power Search from Inspyder Software Inc. is a powerful Windows application for website searching and scaping. Power Search can find or scrape anything embedded in HTML, CSS or JavaScript.

Many websites are built from server-side scripts that dynamically generate content from a database. Power Search is able to crawl through this content and search for any text, even content that's not available from a traditional search engine (such as HTML code snippets).

With Power Search you can crawl password protected websites with ease. Power Search supports both HTTP authentication and HTML based login forms, making it easy for you to search all types of protected pages.

Power Search can also be configured to exclude certain pages from its search, such as message boards, user comments or other pages you might wish to ignore.

Features and Benefits
- Ease of use - Just enter your homepage URL, a word or phrase and start searching
- Choose to include HTML, scripting and style sheet in your search
- Search pages served by your web server, regardless of backend technology (PHP, Perl, ASP, ASP.NET, etc.)
- Construct wildcard searches for partial or fuzzy matches
- Search password protected pages and submit forms
- Export the results to a CSV or XML format file for further analysis or processing.

Power Search requires Windows XP/Vista/7 or Windows Server 2003/2008.

2   Quick Start Guide

When you first run Power Search the default project is automatically loaded but contains no settings. Follow the steps below to configure Power Search for the first time.
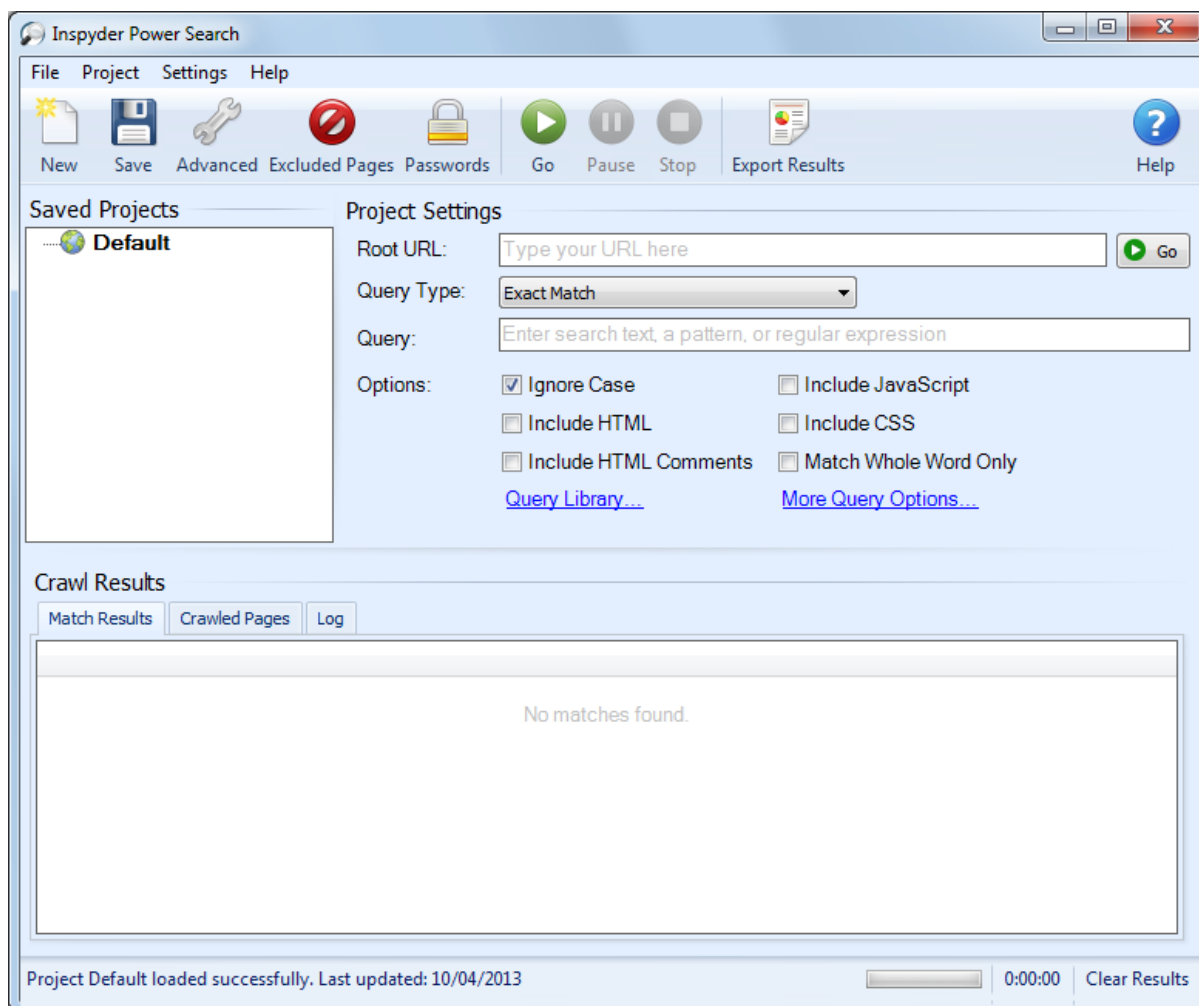
Searching a Website for Text
Step 1   Enter the URL of the website you wish to search in the '**Root URL**' field.
Step 2   Select the "Query Type" from the drop down list. For a simple search use "Exact Match" or use the Query Library to select a predefined query.
Step 3   Enter the text you wish to search for text in the **Query** field.
Step 4   Click the '**Go'** button.
Step 5   Review what Power Search finds in the Crawl Results section.

Scraping Data From a Website
Step 1   Enter the URL of the website you wish to search in the '**Root URL**' field.
Step 2   Change the "**Query Type**" to "**Scrape**".
Step 3   Click the "**Open Scrape Editor...**" button (it will appear when the Query Type is set to Scrape)
Step 4   Use the Scrape Editor to navigate to a page you wish to scrape.
Step 5   Highlight the data that you want to capture and **RIGHT CLICK** to capture it. Repeat for each piece of data you wish to scrape from that page. Note that you only need to train Power Search

on one page, it will automatically find all pages with similar content when it crawls the site.

Step 6 Close the Scrape Editor window.

Step 7 Click the '**Go**' button.

Step 8 Review what Power Search scrapes in the Crawl Results section.
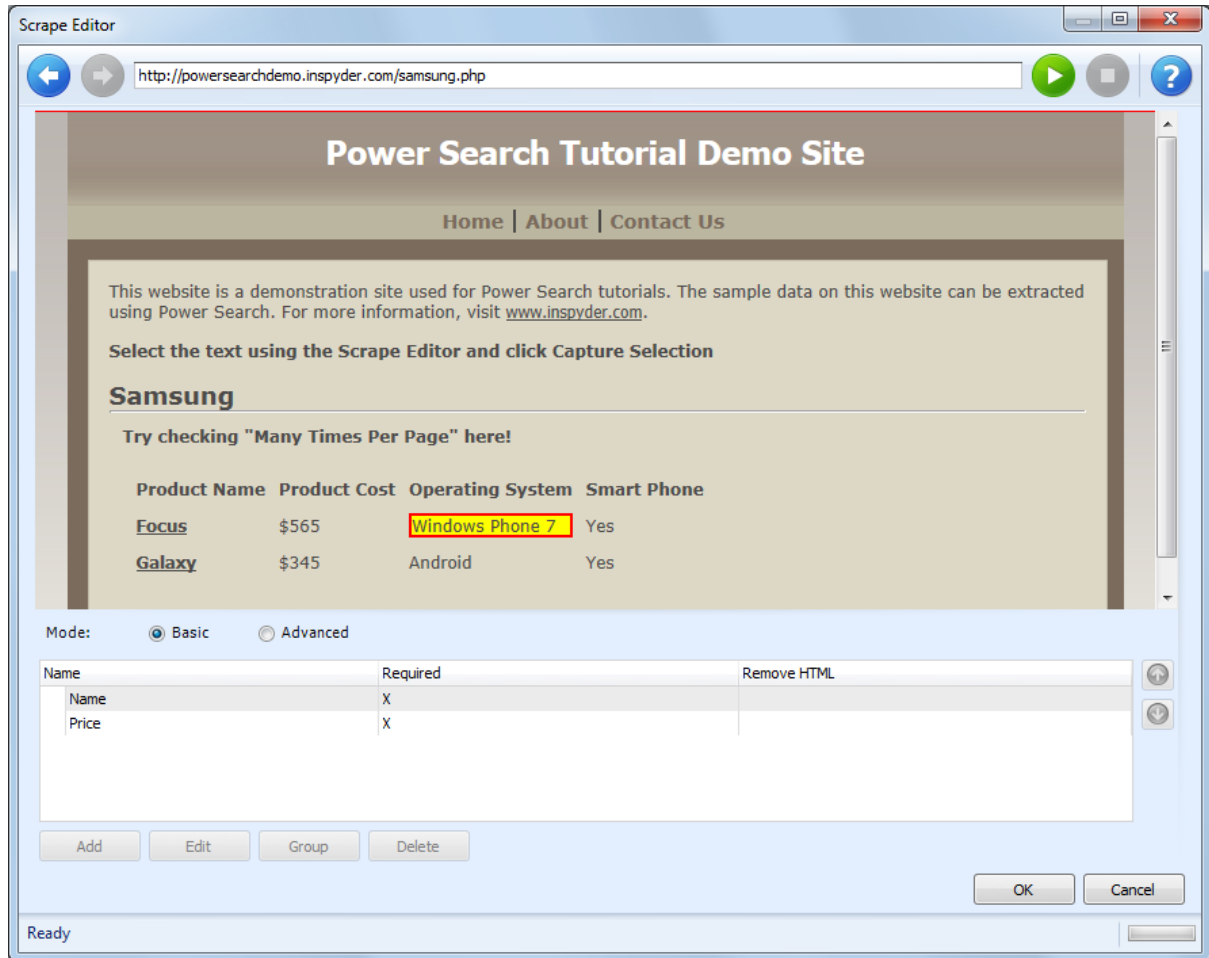
**Hint:** Try modifying the Advanced Project Settings to refine how Power Search crawls the website you are searching. The Advanced Project Settings are accessible from the main menu bar. Click "**Project | Advanced Project Settings**" at the top of the screen.

3   Tool Bar

This section describes the functions of the tool bar shortcuts. These are the most often used features of Power Search and are placed on the tool bar for convenience. Some of this functionality is available by right-clicking on controls in other parts of the user interface, or from the main menu bar at the top of the screen (File, Project, etc.).
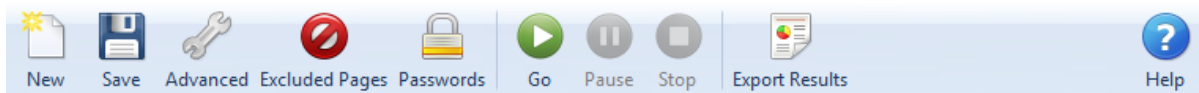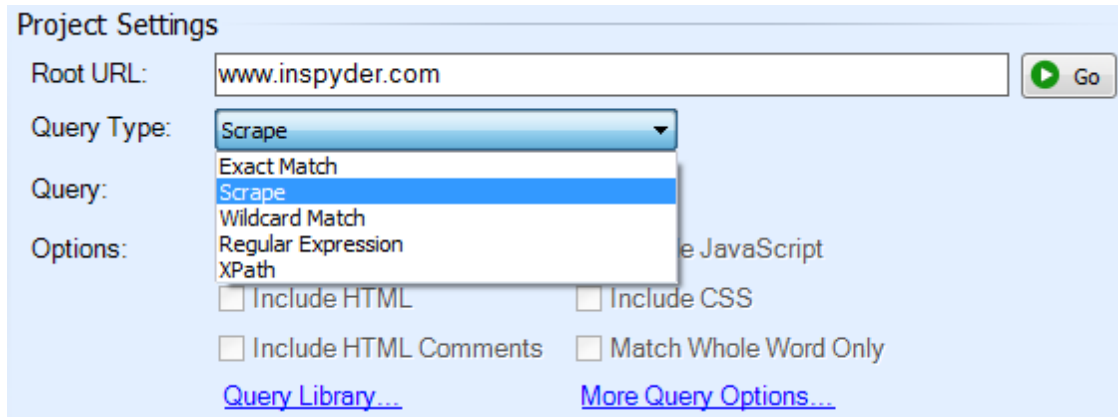
- **New** - The **New Project** button is used to establish a new website that you wish to analyze. You will be prompted for a shortcut name, which will appear in the Projects list, and the root URL of that web site.
- **Save** - The **Save Project** button is used to save the current project settings. 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.
- **Advanced** - This button allows you to access the Advanced Project Settings for the current project. For more information please refer to the Advanced Project Settings section.
- **Excluded Pages** - Opens the Excluded Pages window which can be used to control which pages the crawler visits on a website.
- **Passwords** - Opens the Passwords and Forms window which is used for configuring access to password protected websites.
- **Go** - This button starts the scraping/searching process on the current Root URL. The crawl can be interrupted at any time by clicking on the **Stop** button.
- **Pause -** This button temporarily pauses the crawling process. You cannot close Power Search while paused, but this feature is useful if you wish to Hibernate or put your PC into Standby without aborting your current crawl.
- **Stop -** Use this button to stop the crawling process.
- **Export -** This button provides a way to export the query results. For more information please refer to the Exporting Results section of this document.
- **Help** - Opens the help file.

The main menu contains most of the functionality of the button bar, but extends that capability with a few additional commands. The most notable menu additions are the "**File | Save As**" and "**Settings | Scheduler**" commands.

4   Project

A project contains settings that are typically associated with a single website. This section describes the options, fields and features of the Project Settings area.

Root URL - Specify the address of the homepage for the website to be analyzed. 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 website requires the use of SSL, than you must specify the "https://" prefix.

Query - Enter your Query text here. The Query can be a specific word or phrase, a text scrape, a wildcard match or a Regular Expression. If a Wildcard match or Regular Expression is used, a special format can be used to capture data from the website into the results. For more information, see Query Types.

Go Button - Use this control to begin crawling from the Root URL. This button has the same functionality as the **Go** button on the tool bar. It is included in this location for convenience.

Query Type - The type of text that is entered in the Query field. For more information, see Query Types.

Ignore Case - If this option is checked, the query will not be case sensitive.

Include HTML - If this option is checked, all HTML code will be included in the search. By default this option is unchecked, so only visible text on the website will be searched (HTML markup will be ignored).

Include HTML Comments - If this option is checked then all HTML comments will be searched. This could be used to search for hidden text, such as copyright notices which are often placed inside HTML comments.

Include JavaScript - If this option is checked then JavaScript code will be included in the search (this included inline JavaScript as well as JavaScript source files (*.js).

Include CSS - If this option is checked then CSS code will be included in the search (this includes inline CSS, CSS in "<style>" tags, and CSS source files (*.css).

Match Whole Word Only - If this option is checked Power Search will only return query matches that are surrounded by word boundary characters (such as spaces). For example, of the query were "ton" and Match Whole Word Only were enabled, Power Search would **not** match the word "button". Match Whole Word Only is only available is the Query Type is set to Exact Match.

Query Library - This link provides access to Power Search's list of predefined queries. When selecting a Query from the Query Library, all existing query settings (such as Ignore Case and Include HTML) will be set based on the newly selected query.

More Query Options - This link provides access to additional Query options not listed in the main Project Settings area.

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

> **Windows 2000/XP:**
> C:\Documents and Settings\[username]\Application Data\Inspyder Power Search\
>
> **Windows Vista/7:**
> C:\Users\[username]\appdata\roaming\Inspyder Power Search\

4.1   Query Types

Power Search has four different query types. Each one is used for a slightly different purpose an has different capabilities. They are explained in detail in the following sections. A brief summary is provided here:

Exact Match - Use this mode if you want to search a website to determine which pages contain specific text. This is useful if searching a website for a product name, some HTML code or a filename.

Scrape - This query mode is used to extract structured data from a website using a graphical browser. (Technical users may benefit from knowing that XPath is used in Scrape mode.)

Wildcard - This query type is similar to Exact Match but allows you to match patterns of text using special wildcard syntax. For example, you could use wildcard mode to find all words that begin with "in" and end with "der" such as "Inspyder". Advanced Wildcard queries can also be used to extract data from website in situations where Scrape mode isn't suitable.

Regular Expressions - This query type is for advanced users who need maximum control of their queries. It allows you to search a website using the power of regular expressions.

4.1.1 Exact Match

The Exact Match search type is used to find specific text on webpages. This functionality is typically used to find references to specific items; such a source files, old product names, HTML code snippets, copyright notices, etc.

By utilizing the Negative Match feature, Power Search can also provide a list of pages that DO NOT contain a particular piece of text. For example, you could find all pages on a website that do not include a link to the company privacy policy.

Exact Match will find the entire text as it is entered into the Query field on the website. For more advanced matches see Wildcard Match and Regular Expressions.

4.1.2 Scrape Mode

Scrape Mode provides an easy way to generate a query that will extract data from most websites. This mode is significantly different from the other available query types because it does not rely on a simple

text based query.

Scrape mode relies on capturing one or more text fields from a website. These fields can be generated for you automatically using the Scrape Editor's Point-and-Click interface.

Basic Mode

To begin using Scrape Mode, select **Scrape** as they Query Type, and then click the "**Open Scrape Editor...**" button. (This button will appear where the Query text box was.)
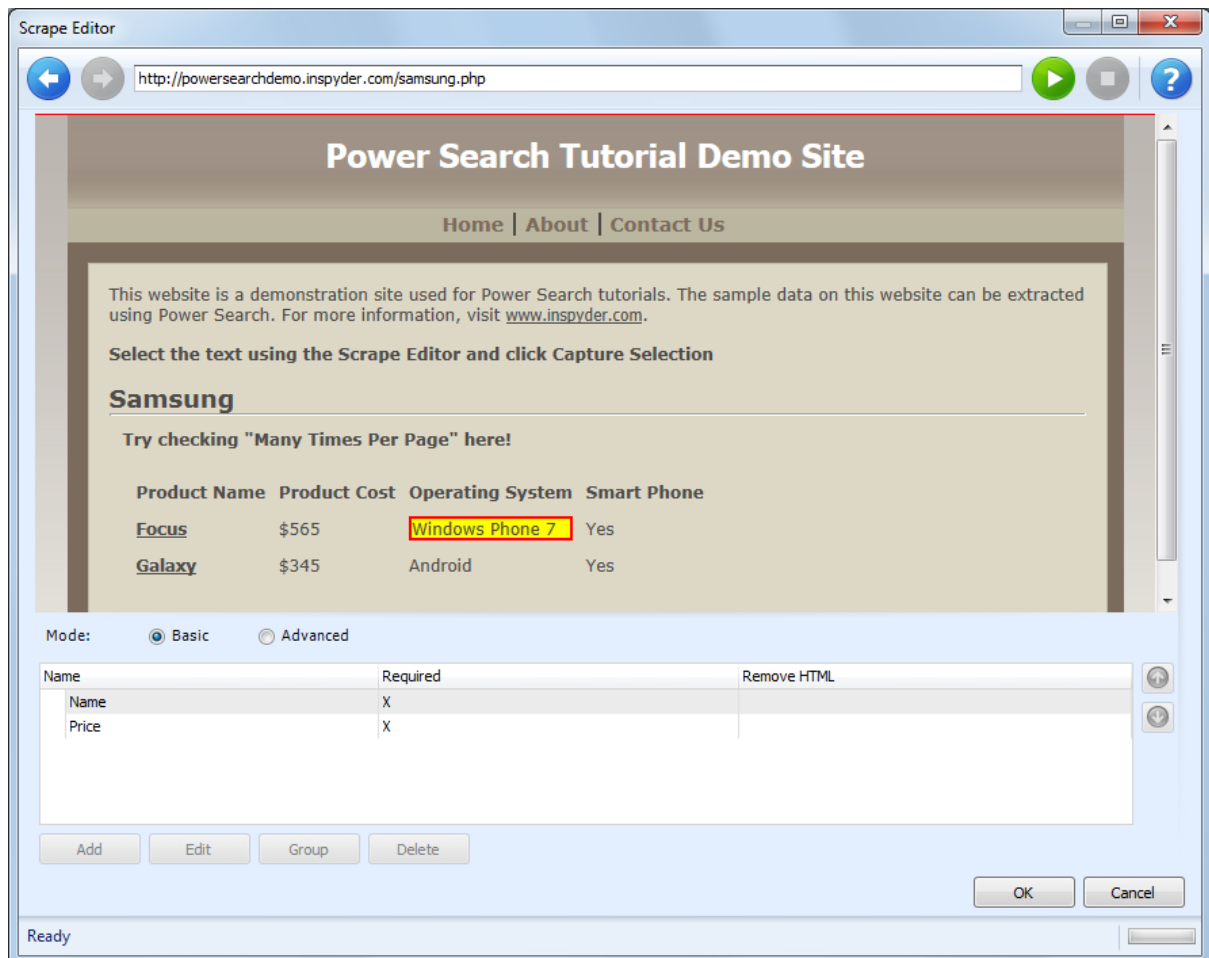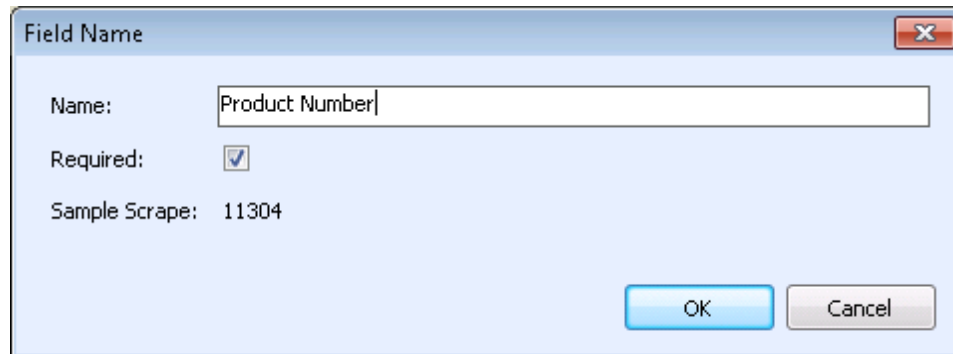
The main element of the Scrape Editor is the embedded browser. It will automatically load the website from your Root URL. (You can maximize the Scrape Editor window to provide a better view of the embedded browser.)

A scrape query is made up of capturing (or saving) one or more fields. To create your first field, use the embedded browser to navigate to a page that contains the text you wish to save. If, for example, you were trying to scrape product names and prices from an online store, you would navigate to a product page that includes the price and product name.

Use the cursor to highlight the text (all or part) that you wish to save and **right-click** the mouse. A window will appear asking for a name for the captured text. The name should reflect what this field contains (such as name, price, description, etc.). The "**Required**" checkbox indicates if this field must appear on each page for it to be considered a match. If the field only appears for some results you want

to scrape, then uncheck it. Click "OK" to store this field name and the associated underlying query.

Repeat the process for each of the fields you wish to scrape from this webpage. Once you have selected all the item you wish to scrape, click OK to close the Scrape Editor window. **Note:** You only need to configure the scrape the fields from one page.Power Search will automatically scrape the other pages on the website to see if they contain the same fields. This means that you only need to train Power Search on **one page** to scrape data from an **entire website**.

If you are trying to scrape data that occurs **multiple times on the same page** you only need to scrape fields for one occurrence of that data. Select the fields you've configured in the list and click the "Group" button. Power Search will automatically group these fields together and look for repetitions on the page.

Tip: You can scrape the URL of an image by highlighting the image and right-clicking as you would for a text field. Power Search knows automatically that this is an image and will capture the URL for the image.

Advanced Mode

Advanced users may benefit from understanding how Power Search works in more depth. Each field that is scraped by Power Search is defined as an XPath query. These queries can be hand written when the Point-and-Click interface won't work. Such cases include writing an XPath query to extract page titles or image "alt" attribute text. To view the XPath editor and fields, set the "**Mode**" to "**Advanced**".

XPath - This field allows you to specify the XPath of text to capture from each webpage. If you have specified a Main XPath query, your XPath here should begin with "./" to denote that it begins relative to the result of the Main XPath.

To remove a field simply select it and click the delete button (trash can). If you wish to edit an existing query, double click on it and it will automatically populate the Name and XPath boxes accordingly. The "Required" checkbox indicates that this match is mandatory. If this XPath query is not found on a page then no result is provided (regardless of wether or not the other queries match).

Mode - Clicking "Advanced" mode will open the HTML document tree to the right of the browser window. You can use this tree to visualize the HTML document structure. Clicking on an element in the tree will highlight that section with a red box in the browser. If you double-click a tree element it's XPath will be copies to the Windows clipboard. (You can then paste that XPath query into the XPath field.)

Text to Remove - Due to the nature of many websites, many fields often come with unwanted prefixes, such as Telephone: 902-232-3930, or Quantity: 2393. The Text to Remove option will tell Power Search to remove any matching text from the Sample Scrape.
For example:
Text to Remove: Quantity:

Original Sample Scrape: Quantity: 20
New Sample Scrape: 20

Regex - Sometimes, an XPath result will contain more than one field. For example, it can contain the Price, Quantity in Stock, and Product number all in one HTML tag. The Regex option tells Power Search to find any match text using regular expression. Regex is often used with the Text to Remove option to help pin point the exact match.
For example:
a <p> tag contains:
Price: $232
Quantity in Stock: 343
Item Number: 379797

If you want to obtain the Quantity in Stock field, you can set:
Text to Remove: Quantity in Stock
Regex: Quantity in Stock: (\d+)

This will tell Power Search to match only the string that has the "Quantity in Stock:" text along with any characters that is a digit and then remove the "Quantity in Stock:" text from that match.
E.g.
Regex will match Quantity in Stock: 343.
Text to Remove will remove "Quantity in Stock:" so you are left with only 343.

XPath Attribute - If you wish to select an attribute from the tag selected by the XPath query, then enter that attribute name in this field.

4.1.3   Wildcard Match

Wildcard searches are a way of easily matching patterns of text. The pattern can be very simple, or very complex. Wildcard support has been included in Power Search to make it easier to find complicated items, or groups of items, on your website.

Basic Wildcards
The basic wildcard character ('*') lets us match zero or more characters in the search pattern. For example; "a*.html" matches:

    a.html
    ab.html
    abc.html

To match a single character, we use the special character "?". For example; "a?.html" will match the following:

    aa.html
    ab.html
    ac.html

It will NOT match words like:

    a.html
    abc.html

Ranges

To match a character within a range (for example, any number character) we can use ranges. A range is encapsulated within square brackets, [], and contains the set. For example, if we wanted to match 'inventoryX.html' where the 'X' is any digit, we could write inventory[0123456789].html. This will match the following:

inventory1.html
inventory2.html
inventory3.html

It will NOT match words like:

inventory12.html

As a shorthand, the following is also valid: inventory[0-9].html.

Multiple ranges within a single set of brackets is acceptable (for example, the letters A to F and the numbers 0 to 9). For example; inventory[A-F0-9].html will match

inventory1.html
inventoryA.html
inventoryB.html
inventoryF.html

It will NOT match words like:

inventoryG.html

It's important to note that a range only matches a single character. To find inventoryXX.html, where "XX" was a number from "00" to "99", the following would have to be used: inventory[0-9][0-9].html. If the search is case sensitive, then [A-Z] and [a-z] are different ranges.

Alternates

Matches that take the form "one or the other" can be specified by using alternates. Alternates are enclosed in curly brackets, "{}", and each alternate is separated by a comma. For example, to find all of the words that start with "ins" and end with "pyder" or "ite" use the following syntax: Ins{pyder,ite}. This will match the following:

Inspyder
InSite

But it will NOT match:

Inspect

Literals

To match a special character as is (such as '*' and '?'), then it must be preceded with a backslash ('\'). For example; 'question\?' will match the following:

question?

But it will NOT match:

---

question!

The same holds true for the '{', '}', '[' and ']' characters.

Extracting Data

In addition to Wildcard matches, Power Search provides a powerful mechanism for extracting data from websites. By including "#DataName#" in your search pattern (where "DataName" is the name of the column you wish to extract data into), any text found in place of "#DataName#" will be scraped and saved.

For example, the title text of a webpage is stored between HTML title tags as follows:
    <title>Inspyder Software Inc.</title>

We could create a Wildcard Match like:
    <title>#WebpageTitle#</title>

(Remember to check "Include HTML") When we run this search on the website, each page's title will be extracted and stored in a column called "WebpageTitle".

It's possible to include multiple fields in the Wildcard Search. For example:
    <IMG*src="#Source#"*alt="#AltText#"*>
This search will extract the "src" and "alt" attributes from all images (where the alt attribute is present). We could enhance this pattern to make the "alt" attribute optional:
    <IMG*src="#Source#"*{alt="#AltText#",}*>
By adding the Alternate syntax ("{alt="#AltText#",}") the match includes "alt="#AltText#" OR nothing" (since nothing follows the comma).

4.1.4   Regular Expression

In addition to Exact Match and Wildcard Match, Power Search has the ability to apply Regular Expressions to the webpages it crawls. The syntax of Regular Expressions is beyond the scope of this manual, but many excellent tutorials are available online, including:
http://www.codeproject.com/KB/dotnet/regextutorial.aspx

To scrape data on a website using a Regular Expression, the expression must include one or more Named Groups. For example, a Regular Expression to match email addresses looks like:
    [a-z0-9\.]*@[a-z0-9]*\.[a-z0-9\.]*
This regular expression will return all pages that contain at least one email address. To capture (or parse) the addresses into a column called "Email", we need to augment the expression as follows:
    (?<Email>[a-z0-9\.]*@[a-z0-9]*\.[a-z0-9\.]*)

The Regular Expression syntax used by Power Search is the Microsoft .NET syntax.

4.1.5   XPath Query

XPath is a query language generally used for querying XML documents. With Power Search it's possible to use an XPath query to select elements from an HTML document. This is one of the most powerful ways to select data from an HTML document.

Enter your XPath query in the Query field. For example, to select each page's title tag, you would use the following:
Query: **/html/head/title**
Attribute: (none)

If you wish to select an attribute from a tag, enter the attribute name into the "XPath Attribute" field to the right of the query text box. (If the XPath attribute is left blank, it's assumed that the text contained in the selected tag should be selected.) For example, to select the "content" attribute of the meta keywords tag, you would use the following query:
Query: **/html/head/meta[@name='keywords']**
Attribute: **content**

If you wish to select multiple attributes at once, you can enter multiple attribute names if they are separated by a semicolon.

All XPath queries are implicitly case-insensitive.

4.2   (More) Query Options

These settings allow you to fine tune how your query is run by Power Search. Note that not all query options are available with all query types. For example, when in scrape mode many features here are unavailable because they do not apply.
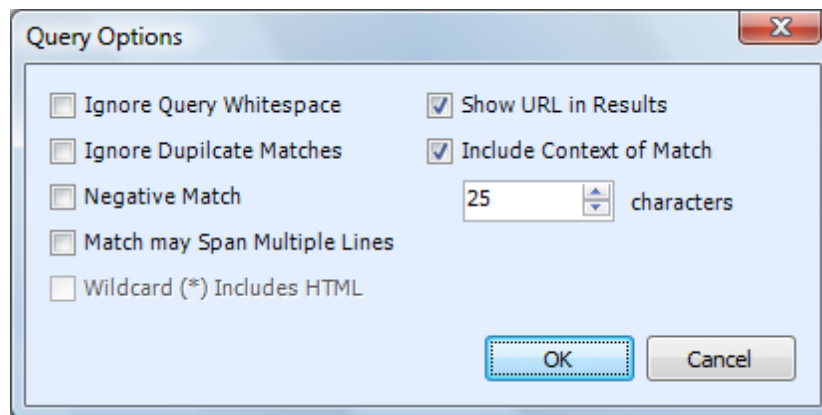
Ignore Query Whitespace - If this option is checked then any whitespace (such as " " or tabs) in the query will be ignored. This is typically used to help make complicated Regular Expressions easier to read.

Ignore Duplicate Matches - If this option is checked, Power Search will only show totally unique results (any result where all columns fields already exist will be hidden). This option is useful if you need to extract a list of unique names, email addresses or some other information and don't wish to manually clean duplicate results from the data.

Negative Match - If this option is checked only pages that do **NOT** contain the query or pattern will be shown. When this option is selected "Include Context of Match" becomes disabled (since a negative match cannot have a context). Show URL in Results will also become disabled (and checked), the URL is the only result returned when using a negative match (essentially providing a list of pages that do not contain the search query).

Match may Span Multiple Lines - If this option is checked then "*" in a Wildcard match will include newline (\n) characters. This is the same as the Regular Expression "Single Line" mode.

Wildcard (*) Includes HTML - If this option is checked, then a "*" in a Wildcard match will include HTML tags. If this option is disabled, the wildcard will only match any text that does not include "<" or ">". This option is disabled by default to prevent Power Search from matching content too aggressively

(which can lead to slow performance and other unexpected results).

Show URL in Results - This option prints the URL that the match occurred on in the output. By default this option is enabled.

Include Context of Match - If this option is enabled the text around the match is shown in a column called "Context". Essentially this shows where on the page the match was found. The length of the context is determined by the number specified in the numeric field (by default, 50 characters).

4.3   Advanced Project Settings

To access the advanced project settings, click the 'Wrench' icon from the button bar, or choose "**Project | Advanced Project Settings**" from the main menu bar.
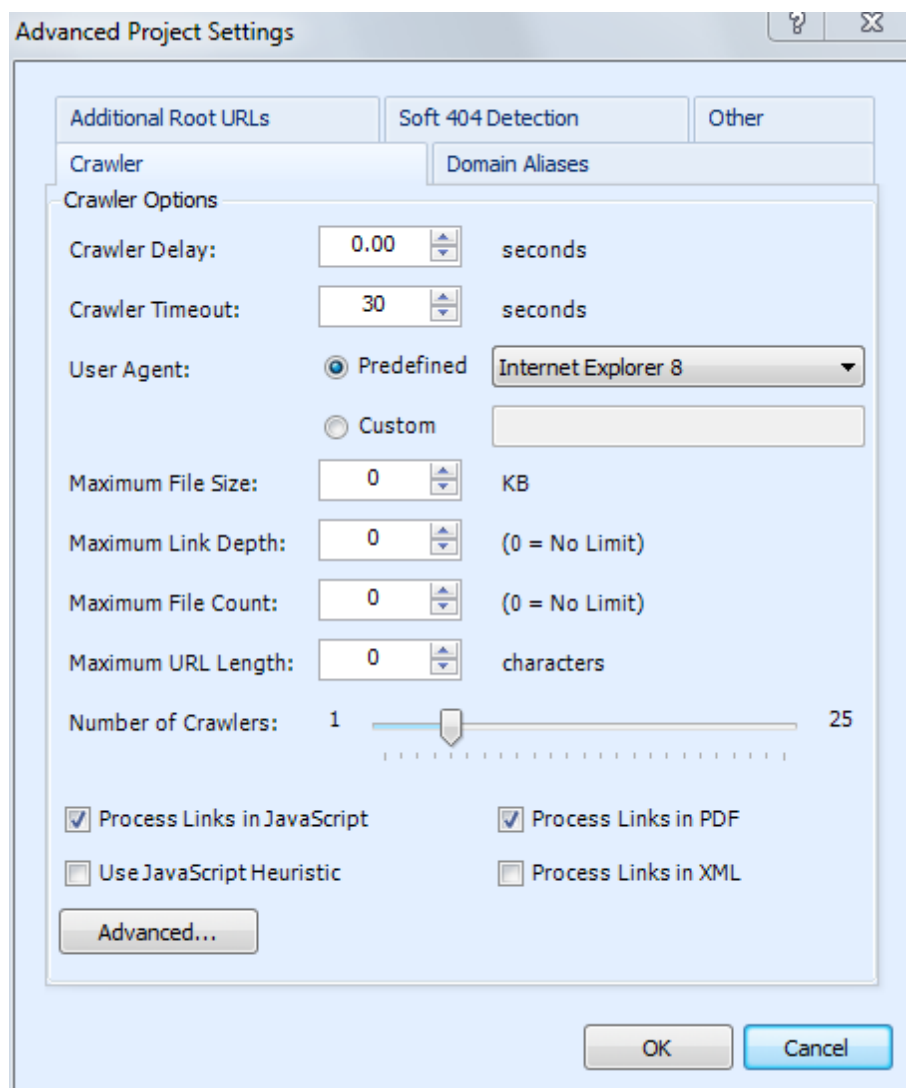
Crawler Options

Crawler Timeout - The Crawler Timeout value specifies the number of seconds that the crawler should wait for a response from your webserver. If your server is slow to respond or has long running scripts, increase this value.

Crawler Delay - The Crawler Delay value specifies how long the crawler should pause between HTTP requests to your server. If Power Search places too high a load on your server, increase this value to slow the crawl down.

User Agent - The User Agent value indicates how Power Search should identify itself to your webserver. By default, Power Search identifies as the Inspyder Crawler. If your site uses browser detection scripts, you may wish to have Power Search masquerade as Internet Explorer or FireFox. The "Custom" option allows you to enter a free form User Agent string. This feature is useful if you have a specific browser your want Power Search to mimic (such as a mobile browser or iPhone).

Maximum File Size - This option limits the amount of data the crawler will download from a single URL. This is useful for websites that have either very large HTML pages (1 MB or more) or script errors that cause pages to generate infinite amounts of HTML. If you receive "Out of Memory" errors while crawling your website with Power Search, try setting a limit of 1024 KB. For most websites setting this value is not required (use 0 for no limit).

Maximum URL Length - This option limits the maximum URL length the crawler will follow. This provides a mechanism to limit the crawler from crawling endlessly on websites that have infinitely recursive links. If you notice that Power Search seems to be crawling similar URLs over and over again, try setting this value to limit the problem. For most websites setting this value is not required (use 0 for no limit).

Maximum Link Depth - This option limits the depth of the crawl for the current project. The depth of the crawl is the number of links (or 'mouse-clicks' for a user browsing your website) from the root URL. For example, if the link depth limit was 1, then all the pages that could be reached in one mouse-click from the root URL would be crawled. If the link depth is 0, then there is no limit.

To limit the crawl by path depth (such as "no pages deeper than 3 subdirectories") you can create an entry in the Exclusion List that looks like: "/*/*/*". This would exclude any URLs that had 3 or more '/' characters.

Maximum File Count - This option limits the crawl by the number of pages crawled. This feature is useful if you are unsure of the website's size and only wish to survey a fixed number of pages.

Crawler Threads - The crawler threads slider is used to adjust the number of simultaneous HTTP requests sent to your webserver. By default, Power Search is set to use 5 threads. Increasing this value can make Power Search crawl your site more quickly, but puts more load on your local computer and on your web server. If you find that your website is running slowly while crawling with many threads, you may need to decrease this value.

Process Links in JavaScript - If this option is checked Power Search will attempt to locate URLs within JavaScript code.

Process Links in PDF - If this option is checked Power Search will attempt to locate URLs within PDF documents.

Process Links in XML - If this option is checked Power Search will attempt to locate URLs within XML documents.

Use JavaScript Heuristic - If this option is checked Power Search will use a more aggressive technique for locating links within JavaScript. If a site makes heavy use of JavaScript (such as for the menu system or other navigation) it may be required to enable this feature.

Advanced Crawler Options

To access the Advanced Crawler Options window, click the "Advanced..." button on the Crawler Options tab page.

Use HTTP Compression - This option tells Power Search to use compression when communicating with your web server. If you see error messages that indicate problem with the "Zip Header", try disabling this feature.

Use HTTP 1.0 - This option tells Power Search to use HTTP 1.0 (rather than 1.1) when communicating with your web server. This should be unchecked unless you experience difficulties crawling a website.

Use HTTP Keep-Alive - This option tells Power Search to keep the HTTP connection (if possible) between multiple requests. Disabling this feature can have a negative impact on performance, but may be necessary with some older servers.

Obey "nofollow" Meta Tag - If checked the crawler will not follow URLs on pages with the no-follow meta tag.

Ignore Protocol Prefix - If checked Power Search will ignore the "http://" or "https://" protocol prefix when determine if a URL has already been crawled. With some websites that mix secure and non-secure sections together is may be necessary to uncheck this option to get a complete crawl.

Ignore HTTP Content-Type Encoding - This tells Power Search to ignore the HTTP "Content-Type" header and use the document encoding (if specified). You should only enable this option if instructed by technical support.

Parse Form "Action" Attribute - By default Power Search does not include URLs found in the "Action" attribute of HTML forms. If this option is checked Power Search will include these URLs when crawling.

Accept Cookies - This option tells the crawler to accept cookies from the website. The default is enabled.

Case Sensitive URLs - This option tells the crawler to treat URLs with different capitalization as different pages. This option should be disabled for most sites.

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.

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.
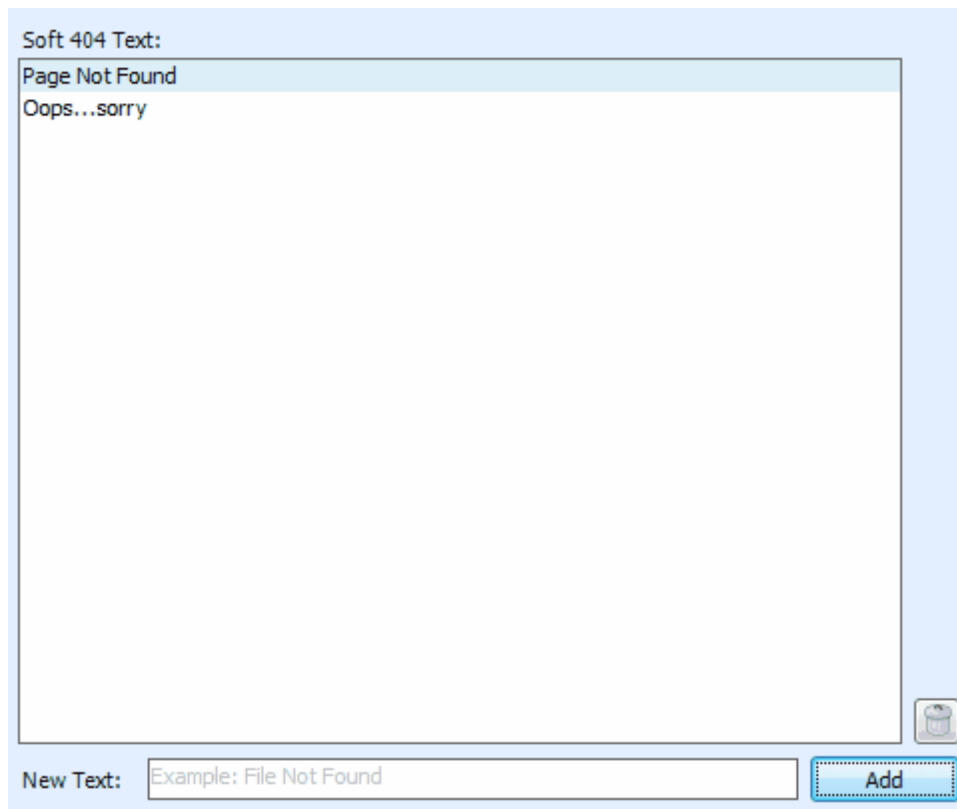
4.4   Soft 404 Detection

Soft 404 Detection allows Power Search to detect which links are broken even if your web server does not correctly return a "404 - Not Found" HTTP header. This is useful if you use a content management system that could report "Missing Pages" but don't necessarily return an HTTP status error. A common example of this is when an invalid product ID number is specified in some online shopping cart systems.

If you find that when crawling your website Power Search crawls lots of incorrect or invalid URLs it's a good indication that you need to configure Soft 404 Detection. Another way to test this is to change your Root URL to point to a page on your website you know does not exist, for example:
http://www.yoursite.com/thispagedoesnotexist
(Replace "yoursite.com" with your actual hostname). If Power Search doesn't detect an error on this bogus Root URL, then you'll need to configure Soft 404 Detection.

To configure Soft 404 Detection just enter some unique text that appears on your website's error page. In the example screenshot below we have configured "Page Not Found" and "Oops...sorry" as our Soft 404 Text. Any time Power Search sees a page with this text on it, it will treat it as an error page. To find out what text your website uses, use a regular web browser to access a URL on your website that you know does not exist.

Soft 404 Text:
Page Not Found
Oops...sorry

New Text: Example: File Not Found   Add

4.5   Passwords and Forms

If your website is password protected or contains content hidden behind a form submission, you can configure Power Search to access that content using the "Passwords and Forms" feature. To access the Passwords & Forms window, click **"Project | Passwords & Forms"** from the main menu bar.

Passwords and Forms allows you to create rules which Power Search will use each time it crawls your website. If a page is encountered with a form (or POST) rule, it will post the specified form data (usually a username and password) to the server when that page is encountered.

Refer to the Excluded Pages section of the manual for more information on how to format the Relative URL field.

The "Method" field tells Power Search what type of credentials to supply to the remote server. There are two methods for supplying login information to a website, POST and HTTP.

Post Method

This is the most common method of logging into a website. If you enter your login information by filling out some type of web based form, then use this method. To configure the login credentials click the '**Form Wizard...**' link under **Actions** to launch the Inspyder mini-browser. Use this browser to log into your website. When you complete the login form, Power Search will ask if you want to save the form data, click 'Yes' and Power Search will automatically create the necessary login rule for you.

Prompt for Manual Login

If you don't want Power Search to store your login credentials, you can enable the '**Prompt for Manual Login**' option. This option only works with websites using cookies for authentication (most websites using form/post based authentication also use cookies). When you click the 'Go' button to start crawling your site Power Search will launch the mini-browser to capture your login session on-the-fly. When you close the mini-browser Power Search will resume crawling with your logged in session.

HTTP Method

The second method is the HTTP method. This is used when the server (rather than the website) is providing the authentication method. If you see a popup login window (with username and password fields) when you access your website with Internet Explorer or FireFox, then your website is using this authentication method. To configure this rule, enter the Relative Path of your login page (if it is your Root URL, just enter "/"), along with the username and password in the corresponding fields (the domain field is optional). Click 'Add' to save the rule.

To remove a rule, select the rule to remove from the list, then click the **"Delete Page"** link under **"Actions"**.

4.6   Excluded Pages

Often it is necessary to exclude some sections of a website from crawling. To do this, Power Search provides the ability to exclude specific URLs or groups of URL from crawling. To access the Excluded Pages window, click "**Project | Edit Excluded Pages**" from the main menu bar.

The Excluded Pages window contains full or relative URLs that should be skipped when crawling. This is important if you have pages or scripts that generate a large amount of redundant content (such as a message board or article pages that are formatted for printing). You might not want these included in the crawl. To add a new URL to the list of Excluded Pages enter the full or relative path in the **Relative Path** field. (As you type the "Actual URL" field will show how your URL is combined with the project's Root URL.) Click the **"Add Rule"** button to save the new URL to the list.

To remove a URL from the list, select it then click **"Delete URL"** (under Actions), or press the "Del" key on your keyboard.
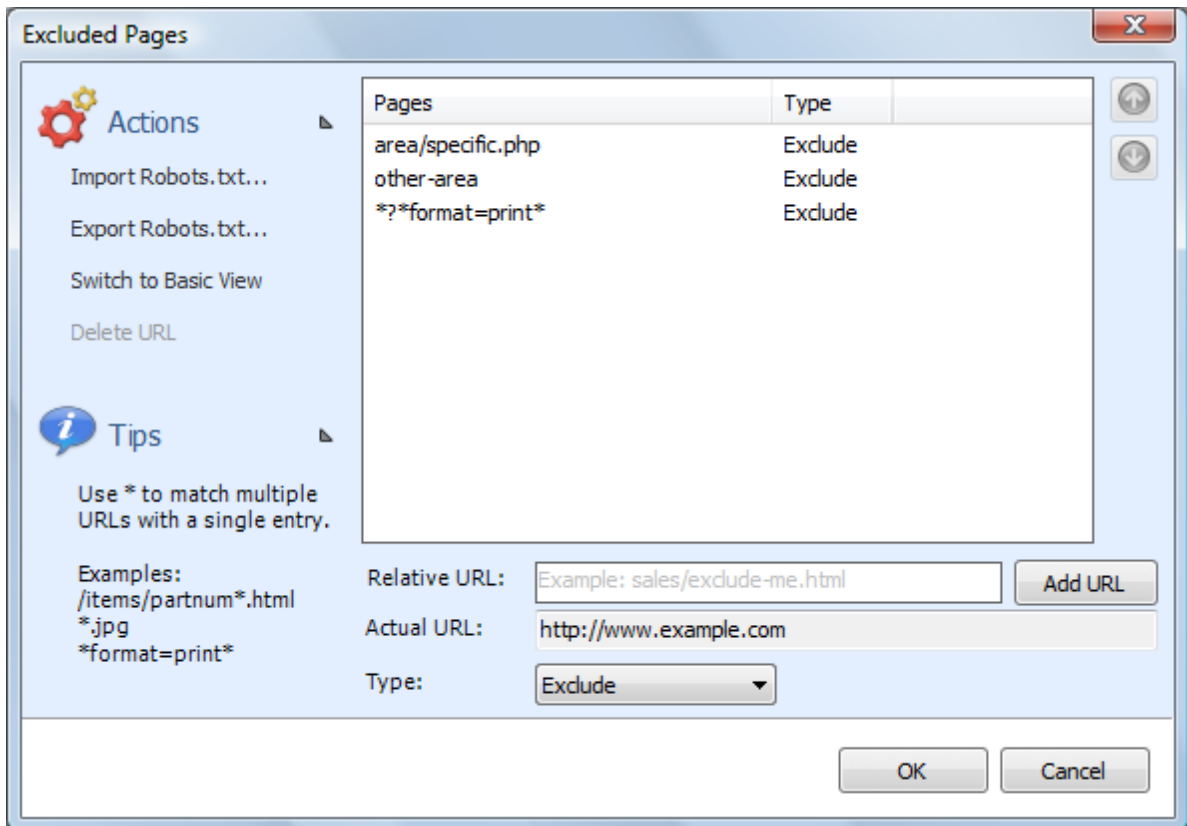
By clicking the "**Switch to Advanced View**" link on the left hand "**Action**" menu you are able to mark a page or section as specifically **included**. The list is matched from top to bottom. If you want to exclude all pages under "www.example.com/content/" except "www.example.com/content/important" than 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.

When a directory (like "/blog") is specified as an Excluded Page, then all of its sub-directories will be excluded (such as: www.example.com/blog/index.html and www.example.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 the Relative Path field.

inventory/warehouse*index   will exclude the following:
                                     inventory/warehouse1index
                                     inventory/warehouse125index
                                     inventory/warehouseindex

userreports/*.txt   will exclude all files with an extension of ".txt"

It is also possible to use the Exclusion List to exclude URLs that contain a particular parameter. For example, to exclude all urls that contain "format=print" in the URL parameter list, you could enter:
        *?*format=print*
This will exclude all URLs that contain 'format=print' somewhere in the URL parameter list (after the '?').

Advanced Tip: The * matches every character including "/". If you want to match any text but **NOT** the path separator ("/"), use an exclamation point in front of the asterisk: "!*". This special case will match any characters **EXCEPT** the path separator.

The "$" can be used to match the end of the URL, like this:

inventory/index.php$ will exclude the following:
                                  inventory/index.php

                                  but will NOT exclude:
                                  inventory/index.php?product=xyz

Internal URLs should be relative to the root URL. URLs beginning with '/' are taken from the domain root (this is recommended). URLs **not** beginning with '/' are relative to your project's Root URL. For example, if the Root URL is "www.example.com/somepage/":

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

As a shortcut, a page can also be added to the exclusion list by right clicking the URL in the Crawl Results.

4.6.1 Importing Robots.txt

Some websites already contain a special file called "robots.txt" which tells crawling software which URLs (or paths) to ignore. Power Search does not automatically obey robots.txt because it is often desirable to crawl the entire website. However, Power Search has the ability to import a robots.txt file into the Excluded Pages list. To access this feature, click **"Import Robots.txt..."** under **"Actions"** on the Excluded Pages window.
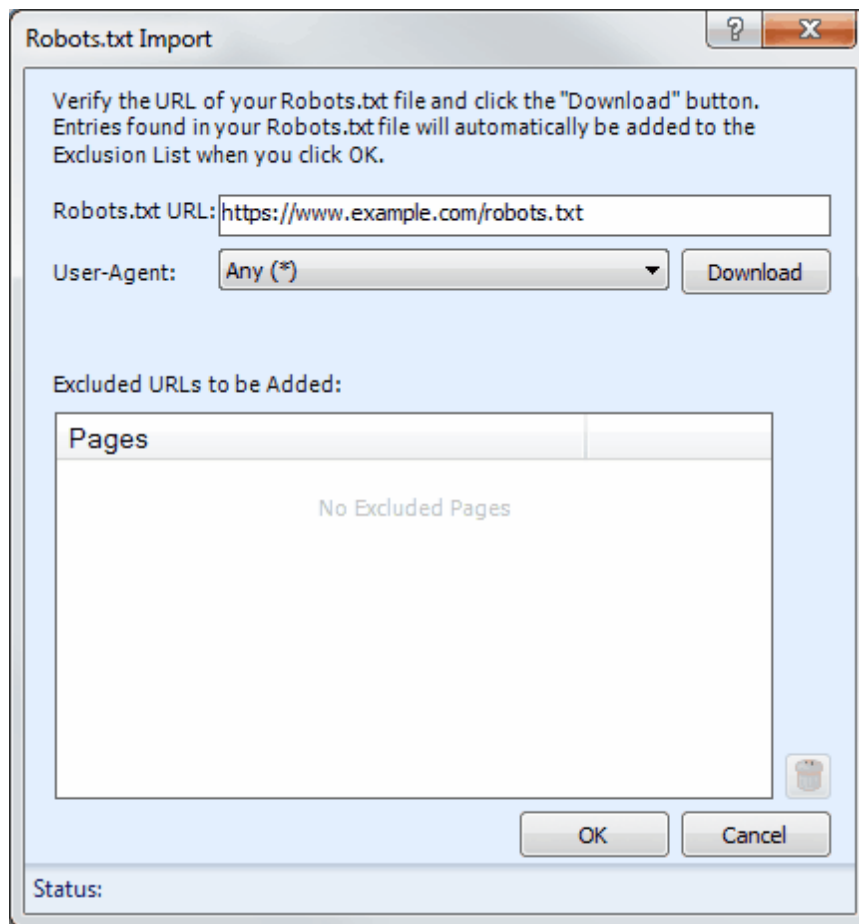
When the Robots.txt Import window opens, it will automatically populate the "Robots.txt URL" field based on your Root URL. You can modify this field if necessary. Click "Download" to fetch and load the file from the website.

The User-Agent field is used to select which rules you wish to import from the robots.txt file. By default the "*" user-agent will be loaded (any bot).

Once Power Search loads the file successfully, the results will be placed in the "Excluded URLs to be Added" list. You can remove any unwanted entries now by highlighting the entry and clicking the "Trash" button (or pressing "Del" on your keyboard). If you click "Cancel" at any time no changes will be made to your Excluded Pages.

When you click "OK", any entries found in the "Excluded URLs to be Added" list will be automatically added to your "Excluded Pages" list.

4.6.2 Exporting Robots.txt

In addition to being able to import an existing robots.txt file, Power Search is also capable of creating (or exporting) it based on the current entries in the Excluded Pages list.

The contents of the Excluded Pages list will be converted into "Allow" and "Disallow" rules and normalized automatically.

The robots.txt file generated by Power Search must be placed at your domain root. For example, if you Root URL was "www.example.com/products/", the robots.txt file created by Power Search must be placed so that it's accessible from "www.example.com/robots.txt" (not in the "products" sub-folder).

The robots.txt file created by Power Search will only have entries under the "*" user-agent field. If you require different rules for different user-agents, you will need to manually edit your robots.txt file in a text editor (such as Notepad).

5   Exporting Results

The export capabilities of Power Search allow you to save your search results to a file in either XML or CSV format.

CSV (Comma Separated Value) format can be opened by most spreadsheet and database applications (for example, Microsoft Excel and Microsoft Access). It is also possible to open a CSV file with a simple text editor such as Notepad.

XML format can be opened by various applications that accept or understand XML markup. This format is better suited for search results that include multi-line matches (such as when capturing large blocks of text from a website).
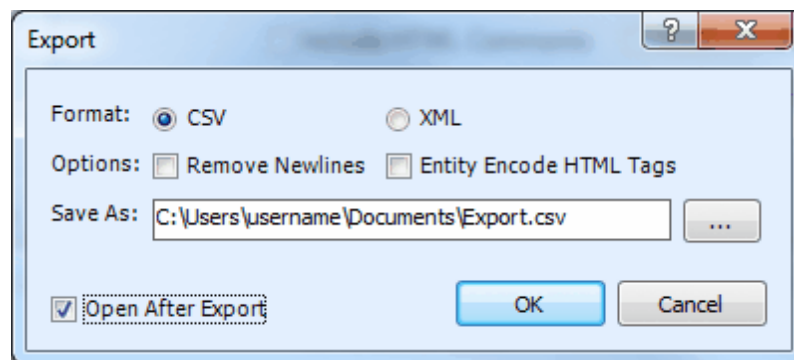
Options

Remove Newline Characters - If this option is checked, newline characters will be removed from the results. This should always be checked unless you specifically require the newline characters in your results. Use this option if you are exporting to CSV.

Entity Encode HTML Tags - If this option is checked Power Search will escape any HTML entities in your export. This should only be checked if your application is importing Power Search's files as HTML content by mistake. (In most cases, this can be safely left unchecked.)

Open After Export - If this option is checked, then Power Search will try to open the exported file with your default application associated with the export format selected (CSV or XML).

6 Project Import and Export

Power Search provides the ability to import and export projects. These features are useful if you need to move projects to a different PC, wish to backup your existing projects, or copy projects between different Inspyder applications.
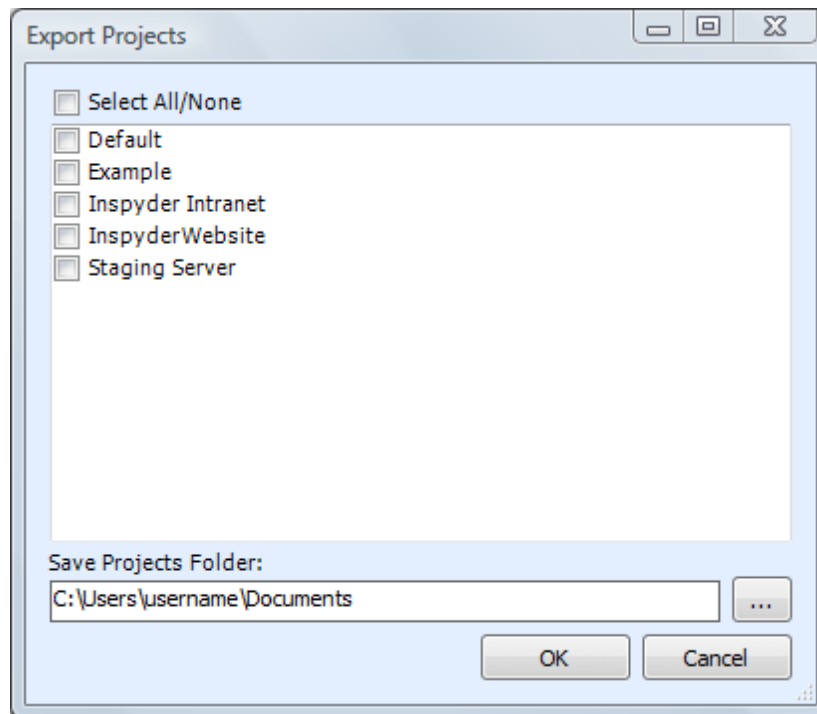
6.1 Exporting Project Files

To export one or more projects, click "**File | Export Project Files...**" from the main menu bar. The Export Projects window will open. Here you can select one or more project files to export and the location where they will be saved.

Each exported project is saved to a unique file. If you chose to export multiple projects, multiple files will be created in the selected output folder. The filenames will have the following format:
"Project Name"."Application".projx

"Project Name" is the name of the project in Power Search. "Application" will be the Inspyder application abbreviation and is used to determine which application generated the project export. All Inspyder project files end with the ".projx" file extension.

Exporting projects from Power Search will not remove them from your list of saved projects. Exporting a project makes a copy of the project file. Any changes made to the project file will not be reflected unless you Import the project back into Power Search.

If you wish to remove a project after exporting it, right click the Project name in the list of Saved Projects and select 'Delete Current Project'.

6.2   Importing Project Files

To Import one or more project files into Power Search, select "**File | Import Project Files...**" from the main menu bar. An Open File window will appear that will enable you to select one or more Inspyder project files.

It is possible to import a project file from an Inspyder application other than Power Search. If you do, only the common elements between applications will be imported (such as Root URL, Excluded Pages, etc.). Application specific settings will be ignored. If you are importing a Power Search project back into Power Search, then all settings will be used.

7   Crawl Results

The Crawl Results section contains 3 tabs: Match Results, Crawled Pages and Log

Match Results

The Match Results tab contains the search results found by Power Search. When searching for an Exact Match, the Match Results will show the URL that contains each match along with the Context of the match. (The Context is where on the page the Query text was found.)

If you are using Wildcard or Regular Expressions, then the Match Results will adjust to include any data fields (or Named Groups) that are specified in the Query text. It is also possible to disable the URL and/or Context when using either of these two Query types.

For information on saving (Exporting) the data from the Match Results tab, see Exporting Results.

Crawled Pages

The Crawled Pages tab contains a list of each page that was crawled and analyzed by Power Search. Only HTML, JavaScript and CSS files are analyzed. (Power Search will analyze any file containing HTML content regardless of the file extension, PHP, ASPX, and other formats are supported.)

Log

The Log tab contains information that can be useful when troubleshooting crawling problems. The log indicates each time a crawl starts and stops, as well as information regarding errors that have occurred. To save the log to a file, right click and select "Save Log...".

To clear all results from the Crawl Results section, click the "**Clear Results**" button located in the lower right hand corner of the Power Search window.

8 Query Library

The Query Library provides access to a set of predefined, commonly used Queries. When you select a Query from the list, an English language description of what that query does is presented in the Description field.
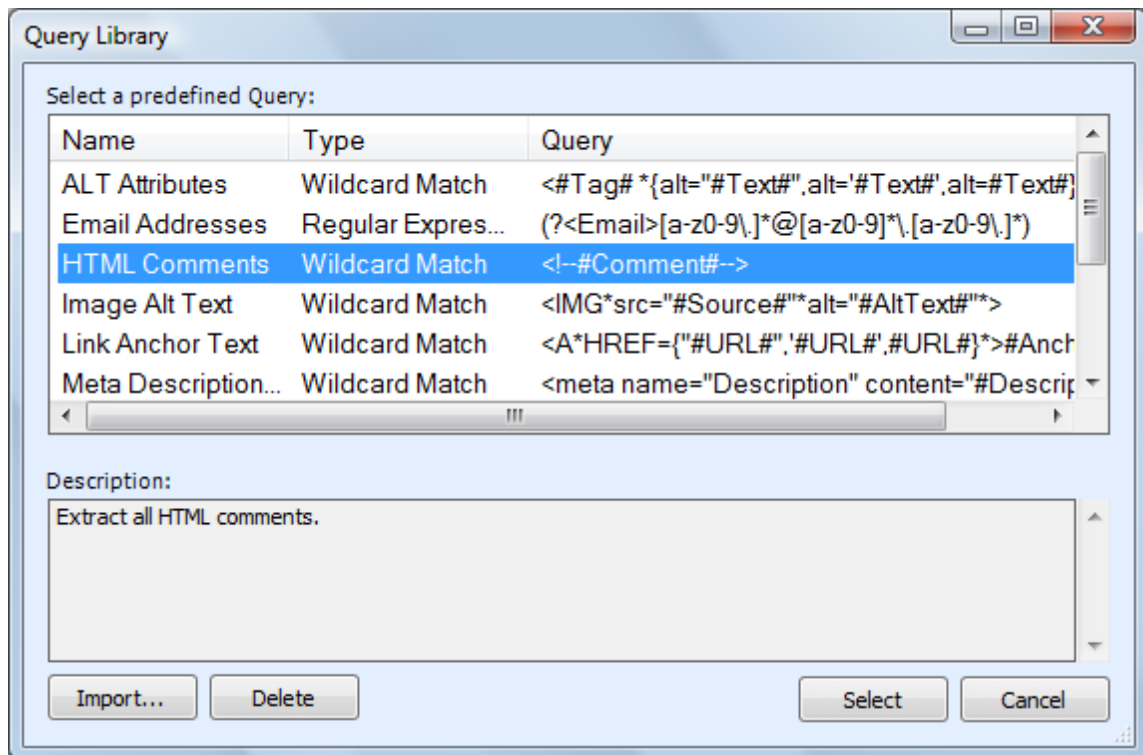
It's possible to add new Queries to the library using the "**Import...**" button. Occasionally Inspyder may publish new queries to our website. You can add these to your library by downloading them, then Importing them with this functionality. It is also possible to import an existing Power Search Project file into the Query library (the non-Query components of the Project will be removed automatically).

To remove a Query from the Query Library, select it and click the **Delete** button.

When you click the "**Select**" button, the settings from the currently selected Query will be applied to your Project. Your Root URL, Excluded Pages and other settings not related to the Query will remain unchanged, however the Query text and any other special settings will be adjusted (such as case sensitivity) for the query to work correctly.

9   Defaut Project Settings

Default Project settings specify the initial configuration used for any newly created projects. Options that are unique to every Project (such as Root URL) cannot be defined as a default setting, however common options (such as Crawler Timeout) can be configured here.

Changes made to the Default Project Settings will also effect **current projects** that use the **same default values**. For example, if you have 10 projects already created in Power Search and you have not changed the "Crawler Delay" setting in any of them (so they are still using the original "0 seconds" value), then those projects are still using the "Crawler Delay" value from the Default Project. Now, if you change the crawler delay in the Default Project from 0 to 1 second, that change will affect all 10 of your current projects.

If you had another project where you had explicitly changed the Crawler Delay, for example, from 0 to 5 seconds, then any change made to the Crawler Delay in the Default Project Settings would not affect this project. This project had the "Crawler Delay" changed within it, so the Default Project Setting value is ignored.

To edit the Default Project Settings click '**Settings | Default Project Settings...**' from the main menu bar. The settings in the Default Project Settings window are nearly identical to those found in **Advanced Project Settings**. For details on what each specific options means, please review the settings in the 'Advanced Project Settings' section of the manual.

10 Scheduler

Power Search has the ability to run in an unattended mode and automatically export the results. To configure Power Search for scheduled operation simply select "**Settings | Scheduler**" from the main menu bar.
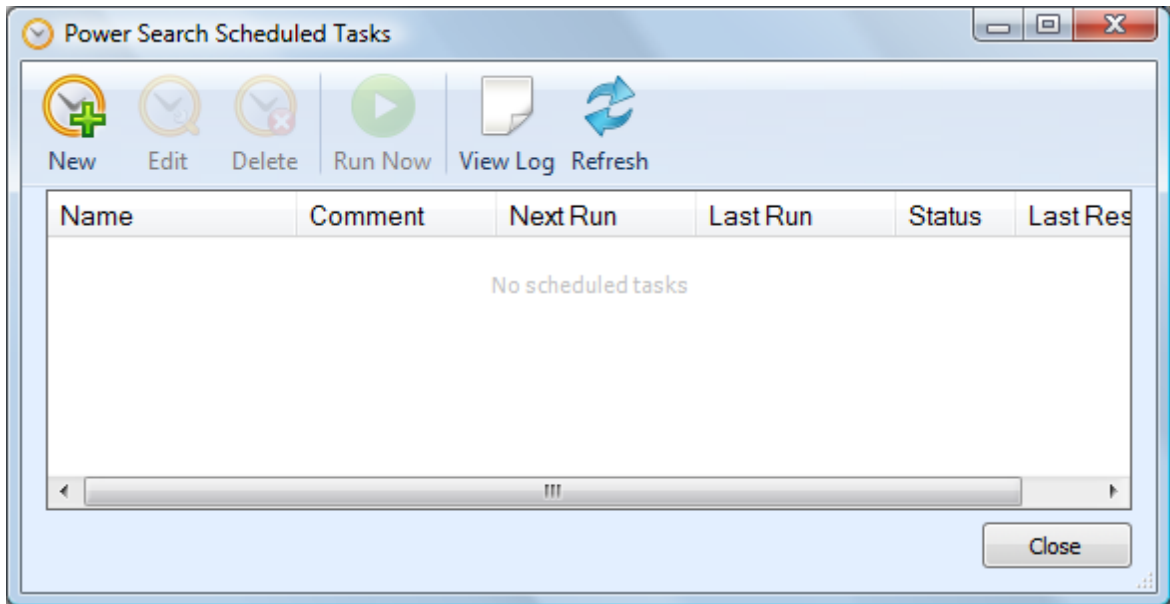
10.1 Scheduling Basics

Scheduling Power Search to run automatically is done by creating one or more tasks in the Power Search Task Scheduler. Each scheduled task contains the settings that will be used when running Power Search automatically. Basic settings include:

- The Power Search Project to load before crawling
- The date and time the task will run next
- How often the task should be re-run

Tasks are run by the Windows Task Scheduler, so Power Search does not need to be open for the scheduled tasks to run, but your computer must be turned on (it will not wakeup automatically).

Tasks configured within Power Search are visible in the Windows Task scheduler. For scheduling options more advanced than those provided by Power Search, please see the Windows Task Scheduler documentation included with Windows.

Name	Comment	Next Run	Last Run	Status	Last Result
Sample Task	Nightly Crawl	1/4/2010 12:00:00 PM	N/A	Ready	1

Any scheduled tasks will appear in the scheduler window. The list of scheduled tasks shows the following information:

Name - The user-friendly name of the scheduled task

Comment - Any comments that were entered when this task was created

Next Run - The next time the task will automatically run

Last Run - The last time the task ran automatically

Status - Indicates if the Task is "Ready" (ready to run), "Running" (is currently running) or "Disabled" (will never run)

Last Result - The result code of the last run. If this number if not zero than an error may have occurred (the Scheduler Log may have more details on any errors that occurred).

10.2   Adding/Editing a Task

To add a new task to the Power Search Task Scheduler, click the "**Add**" button. The scheduling wizard will launch, which will walk you through the steps to creating your new task. When you are finished, click "Close" to save your task to the Windows Task Scheduler. The new task will appear in the list of scheduled tasks.

You can edit a task at any time by highlighting the task item in the list and clicking the "**Edit**" button. This will launch the same scheduling wizard as before with your corresponding settings already populated. Please note, if you have made changes to your scheduled tasks outside of Power Search then those changes will be lost when you edit your task with the scheduling wizard.

It is possible to force a task to run immediately by selecting the Task item, then clicking the "**Run**" button.

Tasks can also be set to "**Disabled**" by right clicking and then un-checking the "**Task Active**" item. When a task is disabled its settings are still saved by the scheduler, but it will not run automatically.

10.3   Deleting a Task

To remove a task from the scheduler, select the task item from the list of tasks and click the '**Delete**' button. This will permanently remove the task from the Windows Task Scheduler, but will not remove any projects, reports or other data associated with that task.

10.4   Scheduler Log

The scheduler log provides a mechanism to review what happened while a scheduled Task was running. To view the log click the '**View Log**' button.
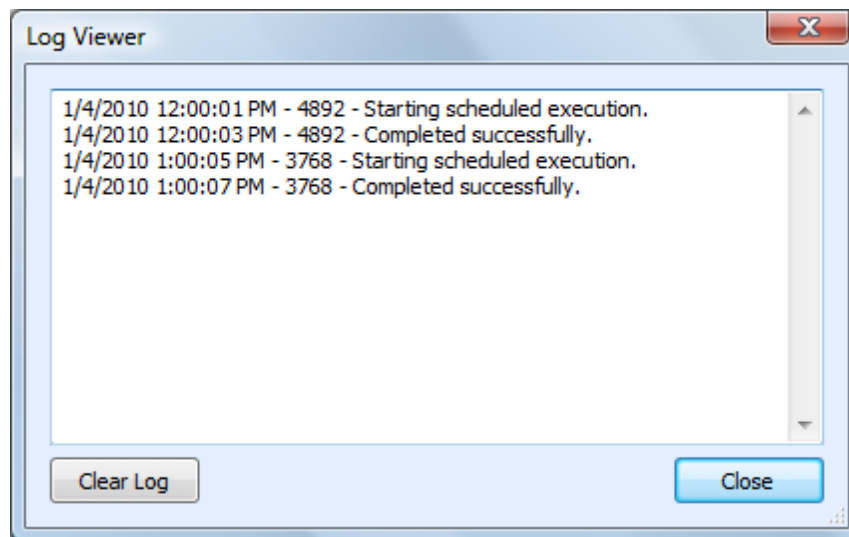
Each line in the log file provides the following information:
- Date/Time the event occurred
- Process ID of the running Task (this is useful for tracing events if you have multiple scheduled tasks running at the same time)
- The log message

To remove all entries from the log, click the "**Clear Log**" button.

10.5   Command Line Interface

If you wish to run Power Search from the Windows command line interface, the following parameters can be used. These parameters are also used by the Task Scheduler when scheduling tasks. If you want to manually adjust your scheduled tasks in the Windows Task Schedule refer to these parameters.

The following are the command line arguments:

-rooturl=	(Optional) Overrides the Root URL configured in the Project
-escape_html	Indicates that the output should have HTML content escaped (for more information, see <u>Exporting Results</u>)
-remove_newlines	Indicates that the output should have newline characters removed (for more information, see <u>Exporting Results</u>)
-output=	Specified the output file for the results (i.e. C:\MyFile.csv)
-format=	Indicates what export format to use (must be XML or CSV)
-project=	The project name to be analyzed (i.e. -project=Inspyder Local)
-email=	(Optional) If specified an email notification will be sent to the following email address (multiple email addresses should be separated with a comma)
-attach	(Optional) Indicates that the exported file should be attached to the email notification
-zip	(Optional) Indicates that the email attachment should be "zip" compressed (this option does nothing unless "-attach" is specified)

Note that if the command line argument's parameter contains one or more spaces, then the enter argument must be enclosed in quotes. (i.e. "-output=My Report.csv")

10.6   Result Codes

These are the codes returned to the Task Scheduler after the Power Search task has been executed. You can view them in the Scheduler window to help troubleshoot any problems.
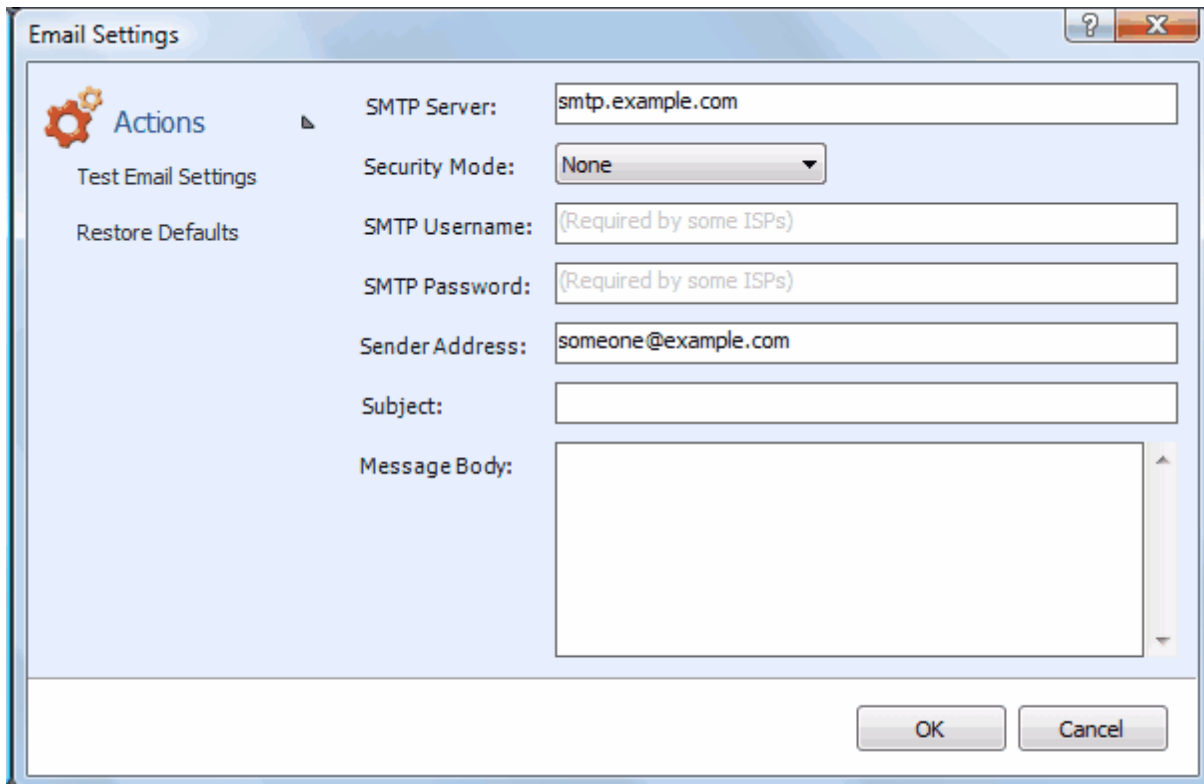
0   Success (no errors)
1   Not registered
        Solution: Run the program interactively and enter your registration information when prompted.
2   Export Error
        Solution: Try running the project and exporting in interactive mode. Check for file in-use, disk full and other possible errors.
3   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.).

11 Email Settings

You can customize the emails that Power Search sends when scheduled crawling has completed. The customization screen is accessed by clicking on the "**Settings | Email Settings**" menu item.

11.1 Server Settings

Configuring your email server settings (SMTP) allows Power Search to send automated emails. To access your mail server settings, click "**Settings | Email Settings...**" from the main menu bar.

Settings

SMTP Server - This is the hostname of your outgoing (SMTP) mail server. If you are unsure what to use, check with your Internet service provider (ISP) or your email provider (such as Google if you use Gmail).

If your provider uses a non-standard port (25 by default, or 465 if SSL is enabled) you can append it to the hostname by separating it with a colon (for example: "mail.example.com:1234").

Security Mode - If your provider uses SSL (Implicit SSL) or TLS (Explicit SSL) select that option here. SSL mode uses port 465 by default. "None" and TLS modes us port 25.

SMTP Username and Password - This is your SMTP username and password. Use these fields if your service provider requires SMTP authentication. If no username and password are supplied Power Search will not attempt to use SMTP authentication.

Sender Address - This is the email address that Power Search will use as the "From:" address for

any emails.

Subject and Body - See the Message Body section in the manual for details.

Actions

Test Email Settings - This action tests your configuration by sending an email to the email address specified in the **Sender Address** field. If you receive this message then your email settings are correct. If an error occurs or you do not receive the email, verify your settings and try again.

Restore Defaults - This action resets the **Subject** and **Message Body** fields to their default values.

11.2 Message Body

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 Power Search before the email is sent.

Tag	Value
#URL#	The project Root URL
#PROJECT#	The name of your Power Search project
#URLS_SEARCHED#	The total number of URLs searched on the website
#RESULT_COUNT#	The total number of results found while searching

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

12   Application Settings

The Application Settings window allows you to change application wide settings that do not affect individual projects. You access this window from the main menu, under "Settings | Application Settings...".
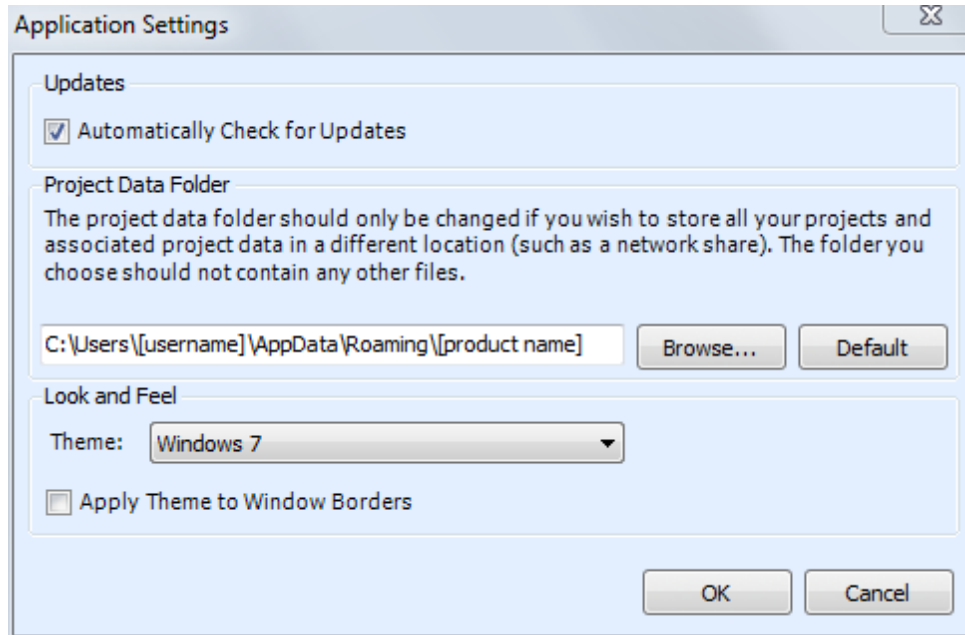
Updates - If this option is enabled, Power Search will automatically check for a new version each time it runs. Unselect this option if you do not wish to see the "A new version is available" message when you start Power Search.

Project Data Folder - This setting allows you to change where Power Search stores it's project and data files. This is useful if you wish to change the default location to a shared network location, or to a folder that is easier for you to backup. Note that you must restart Power Search for any changes to take effect. If you change the Project Data Folder your existing projects will not be moved (you must move them manually). Finally, the folder you select as the Project Data Folder should not be used for anything else. It should only contain projects and files created by Power Search.

Look and Feel - These options allow you change the theme of Power Search. "Apply Theme to Window Borders" allows you to over-ride the Windows border with a style that matches the current theme.

13 About Inspyder Software

Founded in 2004, Inspyder Software is a leading provider of web crawling technologies for content analysis, SEO and website management. These technologies have formed the foundation of our software products and made 'Inspyder' a recognizable brand in the software industry.

Our products are used all over the world by large and small businesses, from independent web designers to Fortune 500 companies. Our user base includes customers in over 50 countries around the world.

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

14   License Agreement

This legal document is an agreement between you, as licensee and Inspyder Software Inc. ("Inspyder"), as licenser. 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), DVD(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 license.

GRANT OF LICENSE: The Licenser grants to You a non-transferable non-exclusive license to use the Product on a single computer. You may physically transfer the Product to another computer provided the Product is used only on a single computer. Installing the Product in a shared environment such as a Terminal Server or otherwise making it accessible to more than the single computer, even if only one user is using the Program at any given time, violates this license agreement. 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 Licenser'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 Power Search to provide an exclusive service such as using Power Search 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 Licenser's entire liability and Your exclusive remedy shall be: (a) the replacement of any media not meeting Licenser's 'Limited Warranty' and which is returned to Licenser 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 Licenser 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 Licenser or any authorized representative of Licenser 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 © 2014 Inspyder Software Inc.

INSPYDER

Inspyder Software Inc.
200-3342 Mainway,
Burlington, Ontario,
Canada, L7M 1A7