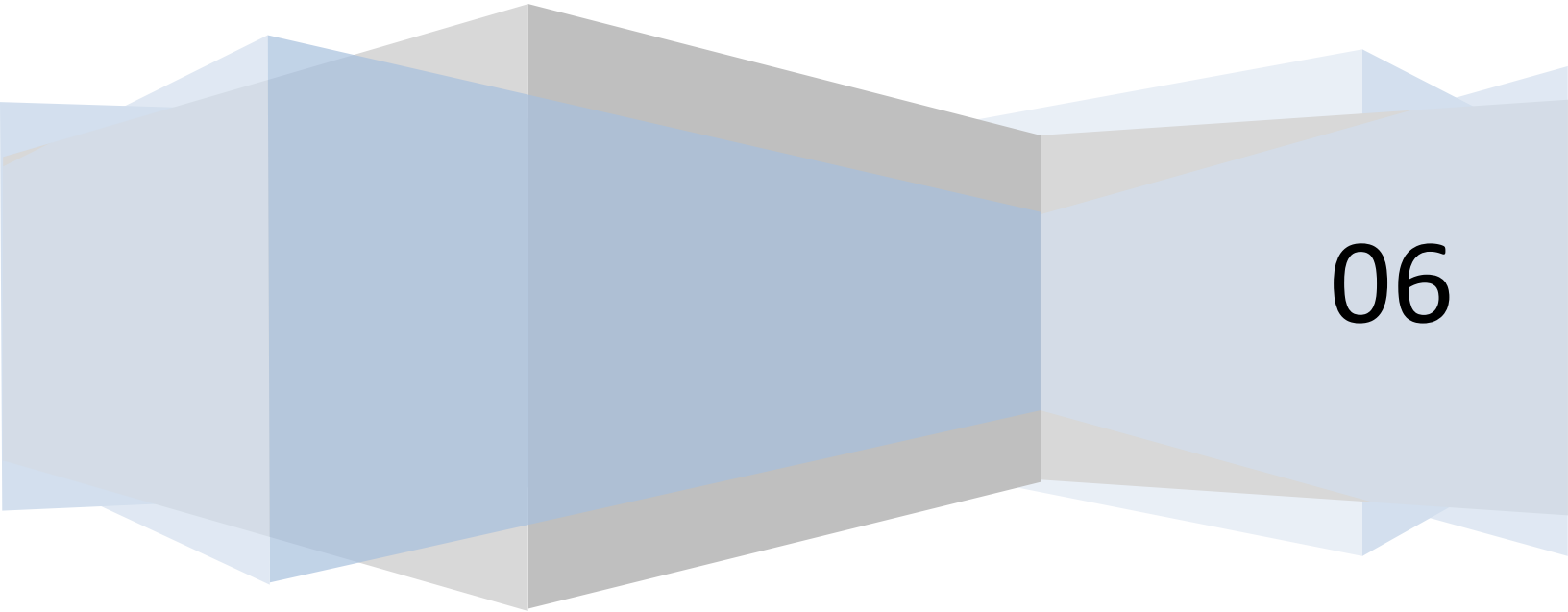
Web Crawler

Crawler User Manual

David Tee



06

Contents

Overview

The crawler is program that allows the user to crawl any (non JavaScript) based websites.

Main Window

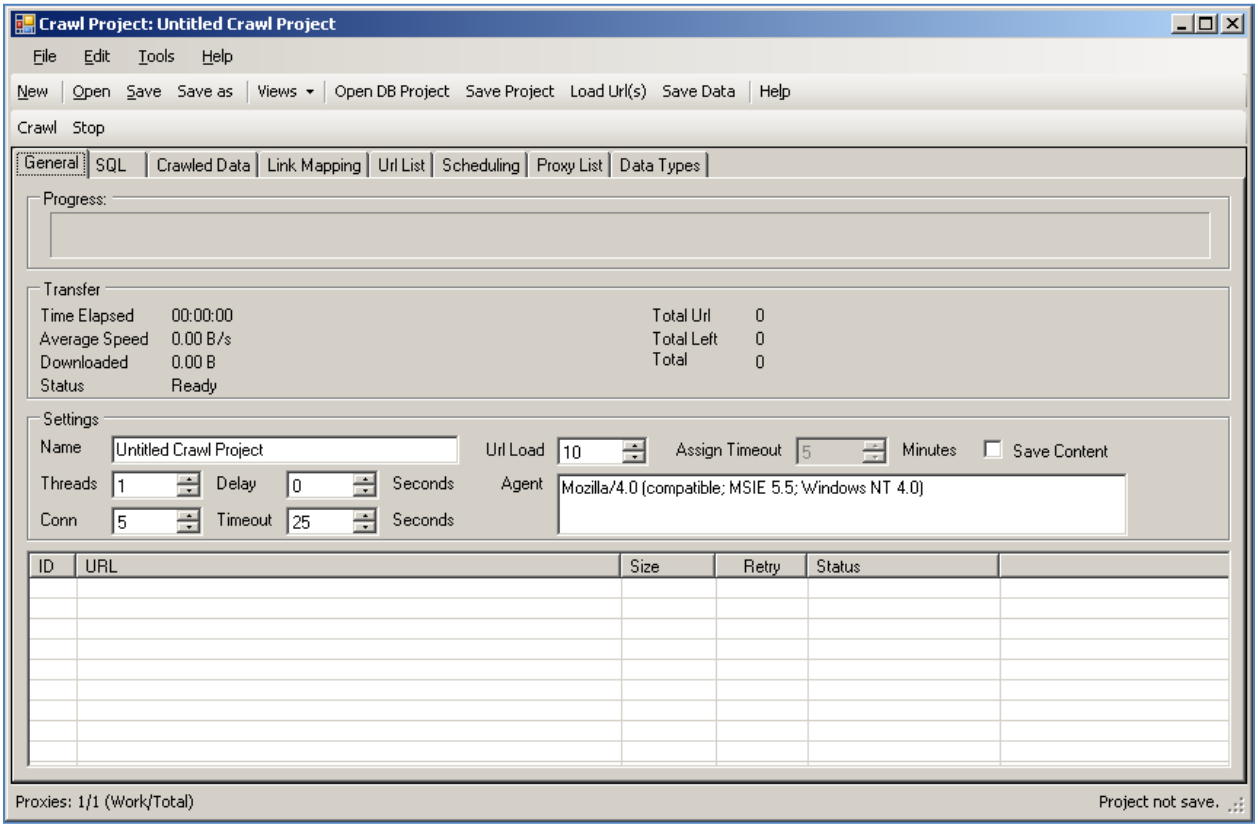


Figure 1: Main Window

This Figure is the main window of what the crawler looks like. In this screen, user can change between tabs to configure a particular crawl project.

Menu

The main includes the follow the items

Menu	Description
File	Contains
New	Create a new Crawl Project
Open	Open an existing crawl project
Save	Save the crawl project to file (or to data base if it a database solution)
Save As.	Save the project to different file
Edit	Contains other Menu Items
Project Information	(Not yet implemented – If Implemented would be the same as the General Tab)
Connection	Change the data base connection for the Scraper to save the crawled data

	or crawl projects.
Tools	Contains Other Menu
Customize	(Not yet implemented – If implemented, would allow the user the change the default values of the fields, such as Agent, Connection, Threads, etc.
Options	Not yet implemented
Help	Contains other Menu Items
Online	(Not yet Implemented) Link to online manual for Web Crawler
Forums	(Not yet Implemented) Link to online forum dedicated to Web Crawler
Regular Expression	(Not yet Implemented) Link to regular expression reference on the website
About	About Window – Shows version number

Tool bar

The tool bar allows the use easier access to the menu item and also some new features. There are 3 different tool bars:

Main Tool bar

Button	Description
New	Create a new Crawl Project
Open	Open an existing crawl project
Save	Save the crawl project to file (or to data base if it a database solution)
Save As.	Save the project to different file
View	Change the Tab group view
Open DB Project	Open a database project
Save Project	Save project to database
Load Url(s)	Refresh the urls
Save Data	Save the project and the data crawled to database.
Help	(Not yet Implemented) Link to online manual for Web Crawler

Crawl Menu

Button	Description
Crawl	Start the crawling process.
Stop	Stop the crawling process.

[Scrape Project] Menu

Button	Description
New	Add a new [Scrape] Project
Wizard	Create a new project using wizard (wizard will generate the tag library – which does not work for most websites.)
Edit	Edit the [Scrape] Project information. (Currently project name, and the sample content for the proxy to search)
Remove	Remove the [Scrape] Project from list. (You may not remove the last [Scrape] Project, a crawl solution requires at least one [scrape] project.

Configuration Tabs

The configuration Tabs allow the user to focus on the specific configuration of the crawler.

General Tab

In this tab, user can set the general crawl solution properties and can also view the progress of the crawl solution. The progress is live and the screen will refresh as the crawl progresses. The screen is divided into a few portions.

Progress

View the overall progress of the crawl. This progress is an estimate and may not be accurate at all as more and more urls are discovered to be crawled.

Transfer

View the data transfer details.

Settings

Property	Description
Name	The name of the crawl solution. This is to distinguish a craw solution from another. (This is necessary if the crawl projects are saved to database.)
Threads	Total Downloads Threads allowed. (This allows the crawler to download web pages concurrently to increase the performance)
Conn (Connections per IP)	This is a feature that's not yet integrated. This feature allow the crawler to limit connection per IP address (per proxy listed).
Delay (seconds)	Download delay in between each download. This allows the crawler to not hammer websites it's crawling. This is a must have feature as some websites limit number of pages downloaded per second.
Timeout (seconds)	Total time given for the downloader to download the website.
Url Load	This is a feature fully integrated that allows the crawler to check out a set amount of url from database. This way, the crawler would check out some amount of url, crawl them and update it back to the database. This allow efficient distributed crawling.
Save context (check box)	This enables the crawler to save the context to the downloaded url to database, into the url data table.
Assign Timeout (minutes)	Assign Timeout allows the database to give a particular crawler a set limit to crawl the given url load. If the urls are not crawled and updated back to data within time, the urls assigned will be assigned to a another crawler that's requesting urls.
Agent	The agent of the crawler that should be send as part of the request string when send to the web server to request a download.

Progress List Box

SQL

This is tab, you can view the SQL code generated to install the database schema. The group box contains one text box that contains the SQL code for Microsoft SQL Server that generates the data table necessary for saving crawl solutions to database.

Sample SQL code from the SQL text box.

```
if exists (select 1 from sysobjects where  id = object_id('Project_Url') and type =
'U')
drop table Project_Url
;
if exists (select 1 from sysobjects where  id = object_id('LinkMapping') and type =
'U')
drop table LinkMapping
;
…
```

Crawled Data

You can view the data crawled during a crawl session. This tab contains a view that allows the user to select tables and the data scraped for the table. (The functionality is the as “Result View Window” in Scraper).

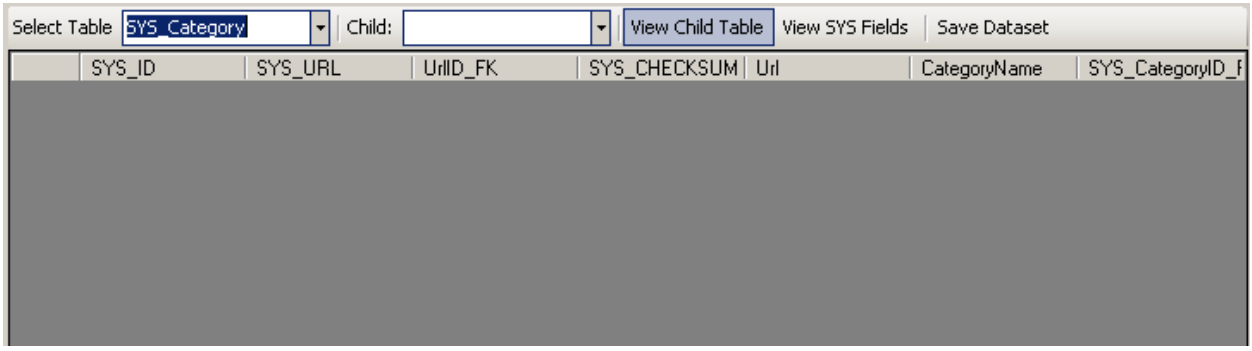


Figure 2: Crawled Data

Link Mapping

Link Mapping is used to configure how the crawler should navigate the website for information. This is where the user can specify the paths (links) tree to take on the website. The crawler will figure out the data relationship using the link mapping configuration.

Url List

In this screen, user can view the urls that are automatically added by the crawling process and that the user manually added. User can manually add a single url or generate a list of urls to be added.

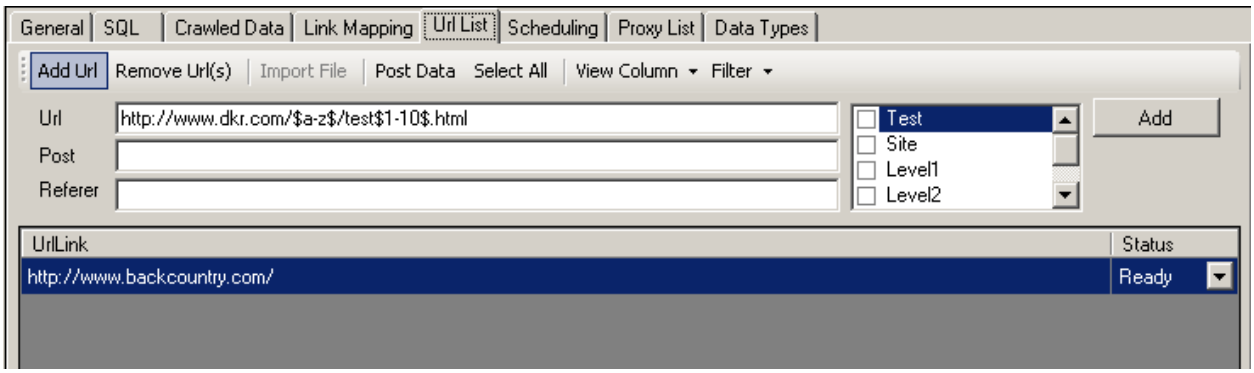


Figure 3: Url List

Scheduling

In this screen, user can define the schedule of when the crawling should place. This is feature to be supported later when the crawler is separated into two parts, Crawl Edit, and Crawl server. The crawl server will use this scheduling configuration to run the crawl project on the given schedule. The schedule is implement using Unix's Cron Job design.

Proxy List

This is feature to be added later. The crawler would support the user of anonymous public proxy to crawl a website. User will be able to add the proxy in this window and the crawler would figure out which proxy to using when download the content of a site.

The use of proxy (if implemented) would only work on websites that do not keep track of user state using cookie or IP. Such sites that keep track of user state would not work with proxies.

Data Types

Data type allow the crawler to type the extracted data to the desired database's data type. This can be viewed as formatting or standardizing the parsed data to the desired data. Data type can be specified in the Tag Library. Each project has their own Data Type.

[Scrape] Project Tab

This tab can be accessed by changing the tab group view (located in the Main Tool Bar – Views button). In this tab, user can edit the configuration for data extraction from a particular webpage. For more information on the Scrape Project, see the Scraper Manual.

Sample Project: BackCountry.com

Back country.com Crawl solution crawls http://www.backountry.com for their category and product information. This project demonstrates how various scraper projects are tied together using Link mapping. Link mapping allows the crawler to figure out which urls to download and scrape the context downloaded with designated the scrape project. To view this project, open "Backcountry.com.sproj".

Here's is the tree of the Link Mapping in the backcountry.com crawl project.

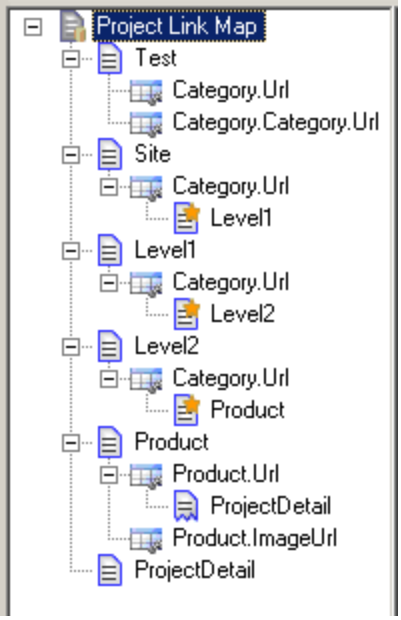


Figure 4: Link Map

Link Map

The link Map in Figure 2 lists all the scrape projects and the url tag nodes that it contains. You can select any of the url node listed in the table to reference the url tag node to another scrape projects (or to itself.)

Url List

Backcountry.com crawl solution starts from one url that serve as the entry point for navigating and crawling the site for its context. If you view the url list, you will see http://www.backcountry.com Url listed there with the status Ready, meaning this url is ready be downloaded and crawled. One of the feature that I should probably add to the UI to view the projects associated with the url (scrape project(s) that will be used scrape the context downloaded).

Sample Crawl Project: Lyrics007.com

Lyrcis007.com crawl solution crawls http://www.lyrics007.com for lyric information and extract the information to Artist and Lyric database. To view this project, open “lyrics007.com.sproj”.