Data Mining Web Sites with TextPipe Pro

Background

This document is a guide to the general principles of harvesting web sites data using TextPipe Pro, our data mining solution.

Why you would data mine a web site -

- Harvest names, addresses and phone numbers of potential clients and competitors
- Extract, cleanse, sort and de-duplicate email address
- Extract and de-duplicate web site URLs from downloaded web sites
- Gather data from your competitor's web sites and then republish it, or use it for sales analysis
- Remove advertising images and HTML from downloaded web sites
- Upload competitor prices into your sales database
- Republish existing text, information or data sheets.

General Principles

There are four key stages to data mining a website

1. Download the site (i.e. obtain a local copy to work with)
2. Throw away unnecessary data
3. Reformat and extract desired data
4. Load into a database

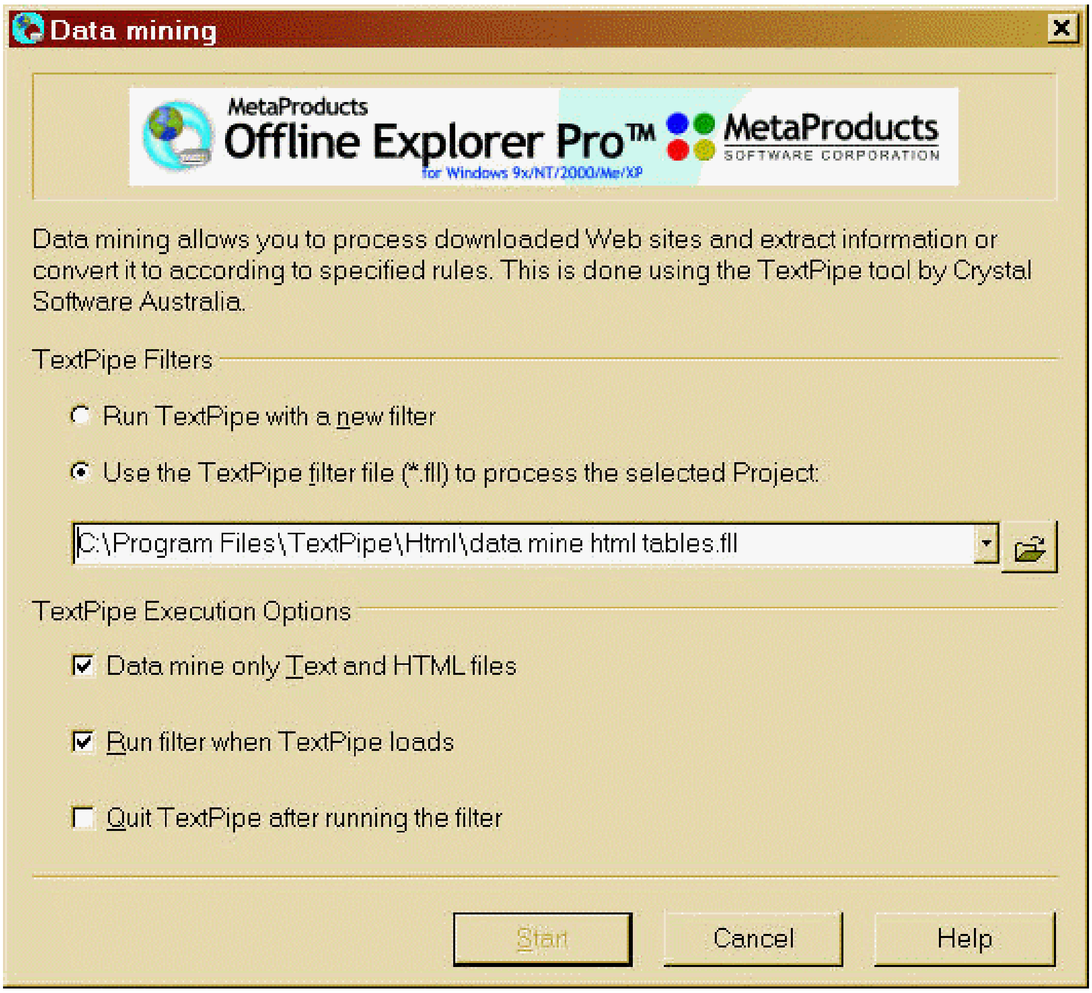
1. Download the Site

The first stage to data mining a web site is to obtain a local copy of it. For this task, we recommend a companion product that integrates with TextPipe Pro, called *Offline Explorer Pro*, from www.metaproducts.com. OE allows you to set the site, control the depth and passwords, set filters on the size and type of file and much more. Sites can be downloaded on a scheduled basis.

When OE finishes downloading a website, it then calls a predefined TextPipe filter to operate on the downloaded data.

We recommend that you set OE to only download text files (.txt, .html, .htm) and ignore all others including image files (.gif, .jpg, .bmp), style sheets (.css), script (.vbs, .js), executable files (.exe, .zip) etc. Doing so will speed up the download time.

When OE detects that TextPipe Pro is installed, it shows a new item under the Tools Menu called Data Mining, which when clicked, shows the following dialog:



Data mining allows you to process downloaded Web sites and extract information or convert it to according to specified rules. This is done using the TextPipe tool by Crystal Software Australia.

TextPipe Filters

○ Run TextPipe with a new filter

● Use the TextPipe filter file (*.fll) to process the selected Project:

C:\Program Files\TextPipe\Html\data mine html tables.fll

TextPipe Execution Options

☑ Data mine only Text and HTML files

☑ Run filter when TextPipe loads

□ Quit TextPipe after running the filter

We have a brief OE Pro tutorial online at www.datamystic.com/textpipe/integration/offlineexplorer.html

2. Throw Away Unnecessary Data

A typical web page consists of many things we don't need - advertising, image tags, formatting, fonts, styles, comments, JScript and VBScript, forms, frames, meta tags etc. So the first stage is to remove all these things.

You can link to our standard filter in web site mining\data mine.fll:

1. Open TextPipe
2. On the File Menu, click Link to filter...
3. Enter C:\Program Files\DataMystic\TextPipe\web site mining\data mine.fll
4. A link to the existing filter is created. You can open the linked filter by clicking the Open button beside it.

It also very helpful for later on, if we simplify tags like

```
<table border="3" padding="3" with="100%">
```

to just

```
<table>
```

This means that we don't have to worry about the exact spacing, order or line breaks between all the attributes of the <table> tag, all we need to know is that there *is* a table tag. You can simplify all the tags to just the tag name (ignoring all the attributes) by following data mine.fll filter with web site mining\simplify tags.fll.

3. Reformat and Extract Desired Data

To convert data from html table format to a CSV (comma-separated value) format that we can easily import into Excel, we link to the filter web site mining\data mine html tables.fll.

Now that your filter is ready, you can test it out on a sample page.

Drag and drop a sample HTML page onto the TextPipe filter window. Then right click the file in the File Grid, and choose *Copy to Trial Input*. Now that the file has been copied to the Trial Run area, you can adjust the filters until you get the desired results. Click the *Trial Run* button to perform a Trial Run.

It's worth noting that you may need to remove other html tables from headers and footers near your data. This must be done manually, because there is no way the software can determine what is junk data and what is not. To remove a table, in the *Special Menu*, choose *Find and Replace (Find Pattern)*. A new search and replace filter is added, ensure it has a find type of Pattern (perl). The add text like '<table>.*</table>'. This will find a start and end table tag with anything in-between.

3b. An Alternative Approach

If the data is not regular, first tag each item you want to extract, and then throw away everything else.

1. Use one or more patterns to match the data you need. Output (replace) the found data with the 'tag' characters ### at the start of the line. E.g.

   Find EasyPattern: Price: $[capture(1 + digits, '.', 2 digits)]
   Replace with: \r\n###$1\r\n

2. Throw away all lines not starting with ###, ie *Filters\Remove\Remove lines\Remove non-matching lines*, with a Pattern:

   ^###

3. If the fields do not always occur in the same order, instead of ### use ###1, ###2, ###3 etc, and use *Filters\Special\Sort* to sort based on the first 4 characters.

4. Finally, we remove the leading ### characters (or ###1, ###2, ###3 etc) using *Filters\Remove\Columns* to remove columns 1 to 3 (or 4).

Now we have all the data, and we just need to convert it to CSV or XML output. We can do this with one whopper search/replace like this:

```
Find EasyPattern:
[ capture( 0+ not cr or lf), cr, lf,
capture( 0+ not cr or lf), cr, lf,
capture( 0+ not cr or lf), cr, lf,
capture( 0+ not cr or lf), cr, lf,
...
capture( 0+ not cr or lf), cr, lf ]

Replace with:
"$1","$2","$3","$4"..."$36"
or
<row value1="$1" value2="$2" value3="$3" ... value36="$36" />
```

4. Load into a Database

Once you have extracted and manipulated the data into a Comma Separated Value format, you can easily load it into a database or into Excel. If the output file extension is .csv, then double-clicking it will automatically open it in Excel.

To load the data into a database, you need to first prepare a database with a table in the right format. To load the data into a database, you need to first prepare a database with a table in the right format (alternatively, you can use a freeform database such as www.asksam.com to avoid having to define a precise database structure, as well as getting very powerful searching and reporting functions for free).

Then you need to manipulate the data into the form

```
  Insert into tablename (field1,field2,field3...) values (value1,value2,value3...);
```

See the example filters in the database folder.

Feedback and Questions

If you have feedback or questions about this documentation, please contact us.

We can also send you updated sample filters from this article, or sample filters tailored to your data processing needs.

More White Papers and Documentation like This

Available from:

www.datamystic.com/docs

TextPipe Pro Evaluation

You can download a free 30 day trial of TextPipe Pro from

www.datamystic.com/textpipe-wp.exe

You can also access our other downloads from

www.datamystic.com/freetrials.html

Please contact us if you have any questions, difficulties or queries.

Contact Details

DataMystic

5 Bond Street
Mt Waverley
Victoria 3149
Australia

Web site:	www.datamystic.com
Phone:	+61-3 9913-0595
Fax:	+61-3 8610-1234