HTML to Text Converter Help Contents

http://www.tali.com
sales@tali.com
support@tali.com

Introduction

Web authors work with HTML. Even if your source files came from another source, once they have been marked up using HTML, they are no longer viewable without an HTML browser. However, it is often necessary to convert an HTML document back to plain text.

A simple approach would be to simply remove all the HTML markup from a file. This would leave a rather ugly, unformatted text file containing many extra spaces, tabs, and line breaks. A better approach would be to interpret the HTML tags contained in the document, much as an HTML browser does, and create a text file containing some of the formatting from the original. HTML PowerTools' HTML to Text Converter takes this improved approach.

The formatting available in a text file is limited, but HTML to Text Converter maintains many of the important aspects such as headings, titles, bulleted/numbered lists (even multi-level numbered lists), paragraph breaks (differentiating between
 and <P> tags), horizontal rules (<HR> tags), and more. Of course, all spaces, tabs, and line breaks used to format the HTML source are dealt with intelligently to eliminate extraneous spaces and line breaks in the converted text file.

Unlike other approaches for converting files, HTML to Text Converter saves time and effort by running in batch mode, converting any number of HTML files with one button click.

Customizable options allow some control of the how the text file is rendered, including choosing word-wrap, setting maximum line length, and choosing to render horizontal rules (<HR> tags) or not. If you ever need to convert HTML documents to text files, HTML PowerTools' HTML to Text Converter will be a very valuable utility in your HTML developer's toolkit.

See also:
Quick Start
Defining a Website Project
Converting Files

Quick Start

Step 1: Select a Website project to convert

1. After starting the program, click Select Project.
2. Click New to define a new Website project to be converted.
3. Enter the required details. Click Help for information on any item.
4. Click OK.
5. Click Select.

Step 2: Perform the conversion

1. Click Convert HTML to Text.
2. Confirm that you wish to convert all the files shown, click Exclude to exclude files, or click Other to select a single file that does not have to be part of the currently selected project.
3. Click Proceed.

Defining a Website Project

Before converting any files, you may define the Website to be processed. You do this by clicking Select Project in the main window, then New in the Project Manager window. After entering the required detailsthat define a project, click OK, then Select.

You can now process all the files contained in your defined project.

You can define any number of separate projects. To select a project to be processed, simply select it in the Project Manager window and click Select.

Note: To convert a single file, that is part of a defined project or not, you can click Other in the Project Window.

Options

All options are set in the Convert HTML to Text Options window, accessible by clicking Options from the main window.

The available options are:

- Render <HR> in Text File - When selected, the <HR> (horizontal rule) tag will insert a string of dashes one character less than the defined line length. When not selected, <HR> tags will only result in a blank line in the text file.

- Convert Tags in PRE Blocks - When selected, HTML tags between <PRE> and </PRE> tags are converted, i.e. treated normally in the conversion. When this option is not selected, HTML tags in PRE blocks will appear in the text file exactly as they did in the original file, without any conversion.

- Insert Blank Line Before List Item - When selected, blank lines will appear items in a list. When not selected, list items will appear without blank lines between them. In other words, when this options is selected, a blank line will be inserted before every <LI> (list item) tag when found between any of the valid list definition tags that support <LI>, namely <OL>, <UL>, <DIR>, and <MENU>.

- Bullet Character/Text - Between zero and five characters to use as bullets for <LI> list items. This will only affect <LI> tags when between <UL>, <DIR>, and <MENU> pairs. (<LI> tags between <OL> pairs are rendered with numbers instead of bullets.)

- Wrap Lines - When selected, carriage returns will be inserted in the file in such a way as to ensure that no line exceeds the number of characters specified in Line Length. When not selected, carriage returns will only be inserted at the ends of paragraphs.

- Line Length - The maximum number of characters allowed per line, when Wrap Lines is checked.

- Convert &character; entities to - these two choices determine how to render character entities for inclusion in the text file. A character entity code always begins with an ampersand & character and finishes with a semicolon ; character. If you do not mind extended ASCII characters in your text file, select Allow Extended ASCII Characters. This will, for example, convert © to the one-character copyright symbol. If you want to ensure that your text file only includes the universal ASCII characters, select Use Simple ASCII Only. This will, for example, convert © to the three-character string (C). You can customize the renderings of these characters using the HTML Rulebase Editor.

- Always Confirm File List - Each time you run the program or select a new project, the file list to be converted (based on the current project's definition) must be refreshed (either by clicking Refresh File List or by clicking Convert HTML to Text). If this option is selected, you will be asked if you want to refresh to file list every time you click the Convert HTML to Text button. If it is not selected, the file list will be automatically refreshed once, and then not again. If you will be making changes to the files contained in your site in between conversions, select this option. Otherwise, it is less irritating to leave this option unselected.

- Text File Extension - Specify here the file extension to give files generated by this program. Normally, this will be TXT. After converting INDEX.HTM, for example, you will have the text file INDEX.TXT.

- Make Backups of Existing Text Files - If this is selected, then all files modified by the program will be backed up in the directory specified in the project's definition.

- Use Character Entities from Rulebase - Select from the list an HTML Rulebase to use for converting &character; entity codes to regular characters. This program is supplied with a single HTML Rulebase containing only character entity information. If you have another HTML Rulebase with customized character entity information, you can select it here. You can customize character entity information in any HTML Rulebase using the HTML Rulebase Editor.

Converting Files

After defining and selecting a project and setting preferred options, click Convert HTML to Text in the main window.

If the file list has not yet been refreshed, it will be automatically refreshed now. If it has already been refreshed and the option Always Confirm File List is selected, you will be given the chance to refresh it again. You can always manually refresh the file list before a conversion run by clicking Refresh File List in the main window.

After refreshing the file list, you will see the list of selected files about to be processed. At this point you have four choices:

- You may click Proceed to accept the list and begin. (If you clicked the Refresh File List button, the Proceed button will instead be OK).
- You may choose to exclude certain files from the current run. To do this, click Exclude Files and use the four Include/Exclude buttons to make your selection. Note that this selection will only affect the current run and will not be saved. When done, click Proceed.
- You may change the permanent definition of the project. To do this, click Modify Project and make any desired changes. Note that this is the same as clicking Properties in the Project Manager window and that all changes are saved. When done, click Proceed.
- You may disregard the currently defined project and select a single file to convert. To do this, click the Other button and select a file. When done, click Proceed. Note that this has no impact on the project definition, and is a one-time selection.

All selected files will now be converted.

Other HTML PowerTools

The HTML tools listed on this page are available from Talicom(R). All run in the Windows environment. Please visit our home page at www.tali.com for more information.

HTML PowerAnalyzer

HTML PowerAnalyzer is a sophisticated tool employing powerful algorithms to scan HTML files and alert the user to all errors contained within them. In addition, a comprehensive report is generated containing a wealth of useful information about each file, and the entire Website.

In addition to all types of HTML syntax errors, HTML PowerAnalyzer will catch invalid &character; entity codes, non-text characters, missing/invalid link references (i.e., files pointed to by HREF, SRC, etc.), missing anchors, and link references containing capital letters (which may be cause problems on case-sensitive Unix servers).

In addition, HTML PowerAnalyzer builds a list of all files included in the project directories that are not included in the Web project and that are not referenced by any files in the Web project. This helps you weed out old and obsolete files that may still be taking up space unnecessarily.

HTML PowerAnalyzer supports the very latest HTML 4, Netscape 4 extensions, and Microsoft Internet Explorer 4 extensions, and can be completely customized. The user can even select which browser (or HTML standard) to analyze for: any proprietary HTML tags (or parameters within standard tags) not supported by the selected browser will be flagged.

HTML PowerAnalyzer's algorithms utilize databases containing all rules of the HTML language -- an HTML Rulebase. The HTML Rulebase Editor allows the advanced HTML user to freely modify any and all aspects of the HTML markup language for his particular purposes. This includes adding/deleting HTML tags and tag parameters, redefining language rules for all defined elements, adding/deleting browser definitions, and more. In other words, the user has the ability to totally customize the logic used in the analysis. Another advantage of the HTML Rulebase is that it can be updated by downloading up-to-date files from Talicom's website, www.tali.com, as they become available. Thus, no matter how quickly the HTML language evolves, HTML PowerAnalyzer will never become obsolete.

In today's rapidly changing WWW landscape, it is not enough to visually check a Website in one or two browsers -- you want to be certain that your markup is perfect and error-free. You also want to know for sure that every single image and hyperlink is perfect, without having to scour your site and test-click every link. HTML PowerAnalyzer will automatically provide you with the certainty you need -- with the click of a button.

HTML Rulebase Editor

A great strength of HTML PowerAnalyzer (as well as some other HTML PowerTools) lies in the customizable HTML Rulebase files that contain the rules of the HTML markup language. Due to the many different implementations of HTML in the real world, and the rapid pace at which the language is presently evolving, it is an absolute necessity to be able to quickly and easily customize any software dealing with HTML.

The HTML Rulebase Editor allows you to do just that. No matter how quickly the vendors of HTML editors and other HTML programs react to changes in the language, they will never

keep up. But you, the user, will always want to be at the forefront. Using the HTML Rulebase Editor, the HTML PowerTools toolkit can always be completely up-to-date.

The HTML Rulebase Editor features a professionally-designed user interface to allow you intuitive and direct access to every relevant attribute of every HTML tag and tag parameter. You can define all aspects of tags and tag parameters for each specific browser (or HTML standard), and even add support for brand new browsers. For example, when Netscape Navigator 5 is released, you can immediately enter all of its new commands and specify them as valid only for that particular browser.

In addition to tag and parameter information, you can modify the lists of defined protocols (e.g., http://, ftp://) and character entity codes (e.g., & ©).

The HTML Rulebase Editor is also an excellent online reference to the entire HTML language. Brief descriptions are included for every tag and tag attribute - and you can add your own.

The flexibility and power that the HTML Rulebase Editor provides for users of HTML PowerAnalyzer are unmatched in the HTML software available on the market today.

HTML Meta Manager

HTML Meta Manager is the fastest, easiest, and cheapest way to guarantee that your Website appears in every major WWW search engine. It allows rapid insertion/editing of Description and Keyword META tags (and TITLE tags) for every page in a Website.

The major WWW search engines, including Alta Vista, Lycos, Infoseek, and WebCrawler, constantly scan the World Wide Web to automatically index every page they find -- including yours. In the absence of any special indicators as to the content of your page, they take a best guess at an accurate description and applicable search keywords. The result is often less than satisfactory, which is why (a) so many searches turn up garbage, and (b) why your site might not come up when someone is searching for it.

So what kind of special indicators can you use to improve the indexing of your Webpages in the search engines? Well, they're called META tags and they can be inserted into every HTML page in a Website. META tags explicitly define a description and keywords for every page in a Website. It is very much in your best interest to include these tags in your pages, if you want to guarantee that your site will come up when a potential visitor is using a search engine. And not just on your home page -- why not have every page in your Website come up separately in a search, improving the chances that someone will click on one of your pages, rather than the competition?

The problem is, adding the required HTML tags to every page in a Website can be a huge job. Some HTML editors (such as Netscape Navigator Gold 3) allow you to define a description and keywords while working on a page, but you still have to manually enter the information for every page separately. The result is that most Websites still do not have the required META tags entered on every page.

HTML Meta Manager is an elegant solution to this problem: it allows you to easily enter a description and keywords for every page in your Website in a single, easy-to-use window. You can enter separate information for each page, or add the same META information to every page in your Website with the click of a button. The program also allows you to easily edit each page TITLE, or to automatically insert the TITLE as the description for every page.

Regardless of the HTML editing environment you work in, HTML Meta Manager can quickly

and easily ensure that all your Web pages are properly listed in the major search engines.

HTML PowerSpell

HTML PowerSpell will allow you to quickly and easily spell-check entire Websites, regardless of what software you used to create them. This program is unique in its comprehensive understanding of HTML files to ensure that you spell-check everything you should while avoiding all the HTML code that you don't want to check.

Unlike the spell-checker that may be built into your HTML editor, you no longer have to laboriously check your work page-by-page. And unlike the spell-checker in your text editor, you no longer have to repeatedly hit Ignore to skip all the HTML elements that you don't want to check.

Advanced spell-checking features that you've come to expect from your software are all included, such as support for custom dictionaries, various options for customizing spell-checking, a context view of misspelled words, and support for multiple languages.

HTML PowerSpell fully understands your HTML files in order to thoroughly and correctly spell-check them. For example:

unbelievable really means unbelievable
resumé really means resumé
<img src="pic.gif" alt="Checkk this text"> should include Check this text
etc.

Typos and misspelled words in your Websites may leave doubt as to your level of professionalism or your attention to detail. Use HTML PowerSpell every time you make changes to your Websites to ensure that your pages are error-free!

HTML PowerSearch

Find and replace utilities, included in editors and word processors as well as stand-alone tools, abound. So why is HTML PowerSearch better?

First of all, HTML files are not text files, even though they are saved as text. HTML files follow a specific set of rules in how their content is read by an HTML browser, and standard find and replace tools do not take this into consideration. For example, in HTML a space, a tab, and a line break are all equivalent. Well-formatted HTML source that is easy to read and work with contains many tabs and line breaks that will never be rendered when the file is viewed in a browser. Your standard find tool will not know to find the search string "hello world" in the following example, yet in HTML it should be found.

> **In this example we have the text "hello**
> **world" separated by a line break and a tab.**

HTML PowerSearch also recognizes &entity; codes and that tags can appear in the middle of search text, so finding "a great day" is easy when the HTML source shows:

> **I'm having a <B>great</B> day.**

Try that with your text editor or word processor! HTML PowerSearch is an HTML-specific tool that knows how to intelligently perform searches on HTML files.

Secondly, performing a search or search & replace across an entire Website using an editor or word processor can be very tedious. HTML PowerSearch handles entire Websites, stored in any number of subdirectories, with one button click.

Thirdly, HTML PowerSearch combines its HTML-specific searching with flexible wildcard searching. You will never be able to return to working on your Websites without the aid of HTML PowerSearch.

HTML Image Scanner

Experienced Web developers know -- and beginners will learn -- the value of using the WIDTH and HEIGHT parameters of the IMG tag: much faster perceived loading of a Web page. When the browser is provided with these parameters, it can set aside a frame for the picture which it will load later, and immediately place all the text on the page.

Unfortunately, inserting these parameters into every IMG tag in a Website is terribly tedious and error-prone. First, you have to use some software to determine the width and height of every image you will use. Next, you have to search for every occurrence of an IMG tag in the site. Then, you have to manually type in the appropriate WIDTH=123 HEIGHT=123. Not only does this process take a very long time, it is irritating. Also, typos during this type of mundane, repetitive work are common, resulting in distorted images and more work.

HTML Image Scanner solves this problem once and for all. With the click of a button it will scan every IMG tag and every referenced image in an entire Website and automatically insert the correct WIDTH and HEIGHT parameters. A number of customizable parameters let you decide, for example, whether to alter an existing parameter (that may intentionally be different from the actual image size) or to leave it alone. You can even enter a list of image filenames which you want HTML Image Scanner to ignore.

Another important attribute in the IMG tag is ALT. This attribute specifies text to be displayed in place of the picture in cases where the picture has not yet been loaded, when the browser is unable to display pictures, when the user has selected not to display pictures, and when Web pages are accessed by the blind. HTML Image Scanner alerts you to every missing ALT attribute, and lets you insert it on-the-fly, complete with the ability to show you the picture then and there.

If you have many images spread across many pages, or if you frequently modify the images included in your pages or add new ones, HTML Image Scanner is an absolute necessity.

HTML to Text Converter

Web authors work with HTML. Even if your source files came from another source, once they have been marked up using HTML, they are no longer viewable without an HTML browser. However, it is often necessary to convert an HTML document back to plain text.

A simple approach would be to simply remove all the HTML markup from a file. This would leave a rather ugly, unformatted text file containing many extra spaces, tabs, and line breaks. A better approach would be to interpret the HTML tags contained in the document, much as an HTML browser does, and create a text file containing some of the formatting from the original. HTML PowerTools' HTML to Text Converter takes this improved approach.

The formatting available in a text file is limited, but HTML to Text Converter maintains many

of the important aspects such as headings, titles, bulleted/numbered lists (even multi-level numbered lists), paragraph breaks (differentiating between
 and <P> tags), horizontal rules (<HR> tags), and more. Of course, all spaces, tabs, and line breaks used to format the HTML source are dealt with intelligently to eliminate extraneous spaces and line breaks in the converted text file.

Unlike other approaches for converting files, HTML to Text Converter saves time and effort by running in batch mode, converting some or all of the files in a Web project with one button click.

Customizable options allow some control of the how the text file is rendered, including choosing word-wrap, setting maximum line length, and choosing to render horizontal rules (<HR> tags) or not. If you ever need to convert HTML documents to text files, HTML PowerTools' HTML to Text Converter will be a very valuable utility in your HTML developer's toolkit.

HTML Date Stamper

Not only is it customary on the Web to include a "last modified on" date in Websites, it is an important indicator to those viewing your pages that the pages are recent and up-to-date. One thing sure to convince a browser not to return to your pages is if they are not updated frequently.

The amount of time and effort required to go into your home page and modify the date is not extreme. However, wouldn't it be beneficial to include a "this site last modified on" date stamp on the bottom of every page in your site? With the Web's new and more powerful search engines (Digital's Alta Vista is a prime example), more and more users will be entering your sites at individual pages and not through your home page. Thus, if you want to show your audience that the site is up-to-date (even if that individual page has not been recently changed), you should have a date stamp on each and every page.

The amount of time and effort to insert the current date into every page, though, is not negligible. HTML PowerTools' HTML Date Stamper will do this for you automatically, on some or all of your Website's pages, with the click of a button.

You provide a simple set of rules that tells the program where to insert the current (or some other) date in any files that you want to be affected. For example, you could define a rule as, "replace all text between the words 'Last modified: ' and the next period with today's date." A number of date and date/time formats are available to choose from.

Project Window

The Project Window, which initially displays all files included the current project, is displayed when clicking Refresh File List from the main window. This window is also displayed automatically when clicking Convert HTML to Text if the file list has not yet been refreshed. (Note: you can skip this window when clicking Convert HTML to Text by holding down Shift or Ctrl when clicking that button.)

If you are satisfied with the file list after refreshing it, simply click OK (to return to the main window, if you clicked Refresh File List to get here) or Proceed.

Alternatively, you have two other options:

- The Modify Project button takes you straight to the Project Properties window to make changes to the current project's definition. When returning to this window, the file list will be refreshed based on your changes to the definition.

- The Exclude Files button allows you to temporarily exclude certain files from processing. After clicking this button, the Project window will contain two lists: one showing all files to be included in the run, and one showing all those excluded. Use the Exclude and Include buttons to move a single file from one list to the other, and use the Exclude All and Include All buttons to rapidly move all files from one list to the other.

  Hint: double-clicking a file name will move it to the other list.

- The Other button provides a dialog box for selecting a single file to be converted. This file may be part of a defined project, but it does not have to be. Note that this has no impact on the project definition, and is a one-time selection.

Main Window

From the main window of HTML to Text Converter you can navigate to the Project Manager and Options windows by clicking the appropriate buttons. You can also begin a conversion run by clicking the Convert HTML to Text button in the center of the window.

The name of the currently selected project is displayed at the top of the window, as well as the number of files currently selected for conversion. If you used the Project window to temporarily exclude any files from the conversion run, the number of files displayed will indicate this fact.

Clicking Refresh File List will re-scan your hard disk to find all files included the project's definition. The resulting list is displayed in the Project window. There, you can temporarily exclude certain files from the conversion, jump directly to the Project Properties window, or select a single file to be converted instead of an entire project.

Note that it is not necessary to click Refresh File List to perform a conversion. Clicking Convert HTML to Text will automatically cause a refresh and provide you with the opportunity to exclude files or select a single file not in the current project.

Hint: if you do not wish to see the Project window before running a conversion, you can hold down Shift or Ctrl when clicking Convert HTML to Text to automatically proceed straight to the conversion itself.

Project Manager Window

This window, accessible by clicking Select Project from the main window, shows a list of all currently defined Website projects. To select one for conversion, click its name and click OK.

To add a new Website project to the list, click New. The New Project window will appear.

To modify an existing project's definition, select it and click Properties. The Project Properties window will appear.

To remove a project from the list, select it and click Remove.

Note that the Cancel button in this window will not undo changes made in the Project Property window.

Project Properties

The Project Properties window appears after clicking New or Properties in the Project Manager window.

After entering all the required details, click OK. To ignore all changes made (and to cancel the addition of a new project), click Cancel.

- Project Name - Enter a descriptive name for the project up to 20 characters in length. You will use this name to identify the project. It will also appear in reports.

- Project Code - Enter a code for the project, up to five characters in length. This code is used to uniquely identify a project.

- Directory - Enter the full path of the project's root directory. Click Browse to browse your hard disk for the correct directory.

- Include Subdirectories - If the project's HTML files occupy subdirectories below the specified directory, make sure this check box is checked.

- Backups to - Enter the full path of a directory to use for backing up files modified by the program. If project files residing in subdirectories of the project's root directory are modified, they will be backed up in corresponding subdirectory names under the backup directory. Tip: to avoid unintentionally modifying backed up files, don't specify a backup directory that is below the project's root directory if you have selected Include Subdirectories.

- HTML File Mask - Here, specify one or more DOS-style file masks for your files. Generally, this will be simply *.htm. You may specify multiple file masks by separating them with semicolons, e.g. *.htm;*.shtml.

- Refresh - Click this button to ensure that your other entries specify the files that you expect. This button is located here for your convenience only, and its use is not required.

- Files in Project Directories - Here, you can review the files included in your project definition. After changing any of the entries in the Project Location frame, the list will be cleared. Click Refresh to re-scan the indicated files.

Convert to Text Window

This window is displayed while a conversion is in progress.

- The Progress frame indicates which file is currently being converted, as well as the progress of the current conversion.
- The Report frame displays the names of the original files and their converted counterparts.
- Click the Pause button at any time to pause the conversion in progress. When in pause mode, this button will become Continue, used for continuing the conversion process (ending pause mode). When in pause mode, you can click Close to abort the conversion.

About HTML to Text Converter

HTML to Text Converter is one of Talicom's HTML PowerTools. To see which version you are using, right-click on the main window's Help button.

Please refer to http://www.tali.com for complete information about Talicom's line of HTML PowerTools for Windows.

HTML Rulebase

An HTML Rulebase is a file containing all relevant facts about the HTML markup language, and is the logical basis for the decisions made by some HTML PowerTools. Rulebase files have an HRB extension and are located in the directory where the HTML PowerTools programs reside.

The contents of an HTML Rulebase can be customized using the HTML Rulebase Editor, available separately. Up-to-date Rulebase files can also be downloaded from www.tali.com.