RESEARCH PAPER | Engineering | Volume : 3 | Issue : 7 | July 2013 | ISSN - 2249-555X

Web Cleaning: Issues and Reviews of PHP Tidy

KEYWORDS

Pooja Mistry
Shrimad Rajchandra Institute of Management & Computer Application (M.C.A) affiliated to Gujarat Technological University

Mr. Jitendra B Upadhyay
Assistant Professor, Shrimad Rajchandra Institute of Management & Computer Application (M.C.A) affiliated to Gujarat Technological University

ABSTRACT *In present years the enhance of the World Wide Web exceeded all opportunity. World Wide Web includes HTML, Web Browser, Web Server etc. Hypertext Mark-up Language (HTML) is the mark-up language for displaying web pages and to display other information in a web browser. When editing HTML page it may happen that an HTML tags are duplicated, unformatted, or empty. It would be good if there is a simple way to fix these mistakes automatically by cleaning and formatting HTML page into nicely layered out markup. This paper will focus on familiarising the new Tidy extension included as part of the PHP5 and onwards versions and how it can be used to make working with and generating properly-formed HTML in a quick and efficient manner. PHP Tidy is a great tool for tidying an HTML page introduced by Dave Ragget.[11] This paper will also focus on issues regarding Tidy and some solutions to these issues.*

1. Introduction

More than 80% of the HTML pages do not conform to the HTML standard, and extra tools such as HTML Tidy.HTML Tidy are needed to mitigate this problem.[3] For badly formatted HTML documents, structure checker programs like HTML Tidy free utility from W3Ccan detect missing and mismatching end tags and map an HTML source into a complete tree.[4] The validated HTML provides better user experience. Tidy is HTML Parser and Beautifier.HTML Tidy library is a project that took Dave Ragget's HTML Tidy program and turned it into set of libraries. Tidy is a new extension for PHP 5 which allows you to parse, validate, manipulate and repair markup documents from within your PHP 5 scripts.[14] It is based on the tidy command line utility released by the W3C, and the extension comes bundled standard with PHP 5 beginning with PHP 5.0 Beta 3.[2]

HTML Tidy helps you to clean up coding errors in HTML and XML files and produce well-formed HTML, XHTML or XML as output.[7][2][9] This research will explore the use of Tidy within your PHP 5 applications. It also works great on the terribly hard to read mark-up generated by specialized HTML editors and conversion tools, and can help you identify where you need to pay further attention on making your pages more accessible to people with disabilities. Tidy is able to fix up a wide range of problems and to bring to your attention things that you need to work on yourself. Each item found is listed with the line number and column so that you can see where the problemlies in your mark-up.

A result of utilizing the HTML::Tidy moduleor W3Cs 'tidy' application does a good job of standardizing the HTML and removing noisy code. However be sure to check that this has not altered your target data in any way and that its effects are consistent across your downloaded dataset. [5]

Tidy uses heuristic rules to translate HTML (well-formed or ill-formed) to well-formed XHTML.[6]

A new version of Tidy is nearing completion which encapsulates Tidy as a library TidyLib, and has been designed for easy integration with other software. TidyLib, like it sounds, is a library version of Dave Ragget's popular HTML Tidy. In fact, one of the motivations for starting the Source Forge project was to refactor HTML Tidy as a callable library. Although the command line tool is great, it is difficult and inefficient to integrate into other software. Tidy is fast, accurate, Flexible and customizable.[12]

PECL is a repository for PHP Extensions, providing a directory of all known extensions and hosting facilities for downloading and development of PHP extensions.[8][13][12] One can also use a postprocessor Like Tidy to Remove formatting, comments and CCSify the Code.[10]

1.1. Examples of TIDY at work

Tidy corrects the markup in a way that matches where possible the observed rendering in popular browsers from Netscape and Microsoft. Here are just a few examples of how TIDY perfects your HTML for you:

1.1.1) Missing or mismatched end tags are detected and corrected

<h1>heading
<h2>subheading</h3>
is mapped to
<h1>heading</h1>
<h2>subheading</h2>[11]

1.1.2) End tags in the wrong order are corrected:

<p>here is a para bold bold italic bold? normal?
is mapped to
<p>here is a para bold bold italic bold? normal?[11]

1.1.3) Fixes problems with heading emphasis

<h1>italic heading</h1>
<p>new paragraph
maps the example to
<h1>italic heading</h1>
<p>new paragraph [11]

1.1.4) Recovers from mixed up tags

<h1>heading</h1>
<p>new paragraph bold text
<p>some more bold text
maps this to
<h1>heading</h1>
<p>new paragraph bold text
<p>some more bold text[11]

1.1.5) Getting the <hr> in the right place:

<h1><hr>heading</h1>

Volume : 3 | Issue : 7 | July 2013 | ISSN - 2249-555X

<h2>sub<hr>heading</h2>
Tidy maps this to
<hr>
<h1>heading</h1>
<h2>sub</h2>
<hr>
<h2>heading</h2>[11]

1.1.6) Adding the missing "/" in end tags for anchors:
<a href="#refs">References<a>
Tidy maps this to
References[11]

1.1.7) Perfecting lists by putting in tags missed out:
<body>
1st list item
2nd list item
is mapped to
<body>

1st list item
2nd list item
[11]

1.1.8) Missing quotes around attribute values are added
Tidy inserts quote marks around all attribute values for you. It can also detect when you have forgotten the closing quote mark, although this is something you will have to fix yourself. [11]

1.1.9) Unknown/Proprietary attributes are reported
Tidy has a comprehensive knowledge of the attributes defined in the HTML 4.0 recommendation fromW3C. This often allows you to spot where you have mistyped an attribute or value.[11]

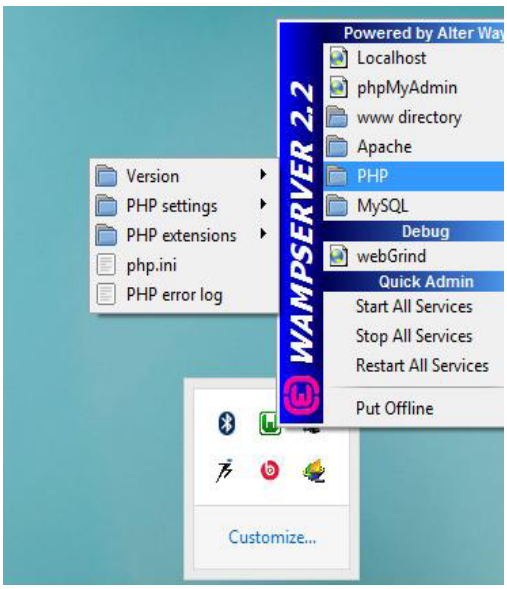
1.1.10) Proprietary elements are recognized and reported as such.
Tidy will even work out which version of HTML you are using and insert the appropriate DOCTYPE element, as per the W3C recommendations.[11]

1.1.11) Tags lacking a terminating '>' are spotted
This is something you then have to fix yourself as Tidy is unsure of where the > should be inserted.[7][11]

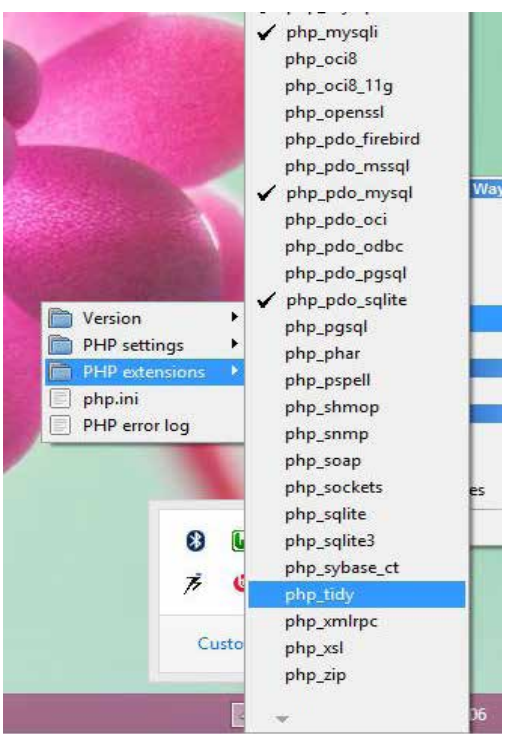
Installation
Although the Tidy extension comes bundled by default in PHP 5.0, it must be enabled in order to be used and requires that the libTidy library be installed on your system. The libTidy.dll file must be placed in your PHP installation folder where all .dll file exists.

PHPTidy with WAMP Server
To use the Tidy extension with PHP one can use WAMP server. To enable the extension of Tidy steps are:

1. Start the WAMP server and start all services.
2. Right click on WAMP select PHP>PHP extension. [In Figure 1]
3. From the menu different PHP extensions are available. From them check php_tidy. [In figure 2]



[Figure 1]



[Figure 2]

An introduction to Syntax
(1)
<?php
$tidy=tidy_parse_file ("http://www.coggeshall.org/");
tidy_clean_repair($tidy);
echotidy_get_output($tidy);
?>[1]

(2)
<?php
$tidy=tidy_parse_file ("http://www.coggeshall.org/");
[1]
$tidy->cleanRepair();
echo $tidy->value;
?>[1]

Volume : 3 | Issue : 7 | July 2013 | ISSN - 2249-555X

(3)
```
<?php
$tidy=tidy_parse_file ("http://www.coggeshall.org/");
$tidy->cleanRepair();
echo $tidy;
?>[1]
```

(4)
```
<?php
$tidy = new tidy();
$tidy->parseFile("http://www.coggeshall.org/");
$tidy->cleanRepair();
echo $tidy;
?>[1]
```

Using Basic Tidy

1.3.1 Parsing Documents

tidy_parse_file($file [, $options [, $encoding [$use_inc_path]]]);[1]

1.3.2 Cleaning and Repairing

tidy_clean_repair($tidy);[1]

1.3.3 Retrieving Output

tidy_get_output($tidy);
which is equivalent to
```
<?php
/* These are equivalent */
echo $tidy;
echo $tidy->value;
?>[1]
```

1.3.4 Dealing with errors

tidy_get_error_buffer($tidy);

Since many common operations in Tidy are based on the above three functions (tidy parse file() , tidy clean repair(), and tidy get output()), the Tidy extension provides two shorthand functions which combine these functions into a single call. These functions are tidy repair file() and tidy repair string() for files and strings respectively.[1]

tidy_repair_file($filename [, $options [, $encoding [, $use_inc_path]]]);[1]

Where each parameter is identical to that found in the tidy parse file() function. Likewise, the syntax for the tidy repair string() function is as follows:

tidy_parse_string($data [, $options [, $encoding]]);[1]

Example of PHP Tidy
```
<?php

$o = array ("clean" => true,

"drop-proprietary-attributes" => true,
"drop-font-tags" => true,
"drop-empty-paras" => true,
"hide-comments" => true,
"join-classes" => true,
"join-styles" => true
);
$tidy = tidy_parse_file("php.html", $o);
tidy_clean_repair($tidy);
echo $tidy;
?>[8]
<?php
ini_set("tidy.default_config",
/path/to/compact_tidy.cfg");
ini_set("tidy.clean_output", 1);
?>[8]
```
Here,
clean=1
drop-proprietary-attributes=1
drop-font-tags=1
drop-empty-paras=1
hide-comments=1
join-classes=1
join-styles=1[8]

Problems with Tidy:

Several people have asked if Tidy could preserve the original layout. But it would be very hard to support due to the way Tidy is implemented. Tidy starts by building a clean parse tree from the source file. The parse tree doesn't contain any information about the original layout. Tidy then pretty prints the parse tree using the current layout options. Trying to preserve the original layout would interact badly with the repair operations needed to build a clean parse tree and considerably complicate the code.[11]

Tidy issues

1. Asian Character Encodings: ISO-1022, ShiftJIS and Big5

Currently Tidy does not transcode ISO-1022, Shift-JIS or Big5 encodings into Unicode. There appears to be consensus that, long term, it can and should. That said, there are several, slightly different mapping tables between Unicode, Shift-JIS and ISO-1022.[11]

2. Escaping <script> and <style> XHTML

There are a number of problems / open issues surrounding the escaping of <script> and <style> tags when producing XHTML output. For those just tuning in, the basic issue is that browser scripts will often contain special XML characters: '&', '<', ']]>' and '<' + '/' + Letter.

The agreed solution is to place <script> source within a CDATA section. This is now done for both <script> and <style> tags. So far, so good. But there are a number open issues and possible unintended consequences.

Conclusion

The objective of this paper was to give an introductory report for cleaning and optimizing HTML/XML using PHP Tidy. The analysis of PHP Tidy shows how PHP Tidy works for cleaning HTML / XML.

However, the above study of paper noted that current version of PHP Tidy still can't perfectly clean the HTML/XML Page Content and having above issues like not support of several encodings, problem with <script> and <style> tag, preserving original document and so on.

Future Work

Regarding to the above problem, the research direction is of fully exploring the new tool which provide all Tidy library functionalities as well as by using Tidy library itself trying to resolve the issues related to Tidy by adding some more features and extending it.

REFERENCE

[1]RechtlicherHinweis, Zusammenfassung (2004), "Tidying up your HTML with PHP" | [2]Hung Viet Nguyen, HoanAnh Nguyen, Tung Thanh Nguyen, Tien N. Nguyen(2011), "Auto-Locating and Fix-Propagating for HTML Validation Errors to PHP Server-side Code" | [3]AykutFirat (2003), "Information Integration Using Contextual Knowledge and Ontology Merging". Cambridge | [4]Boris Chidlovskii (2002), "Information Extraction from Tree Documents by Learning Subtree Delimiters", Website: http://www.isi.edu/info-agents/workshops/ijcai03/papers/chidlovski-subtreeDelimiter03.pdf | [5]James owison, Kevin Crowston(2004), "The perils and pitfalls of mining SourceForge", Syracuse University. | [6] Yelena Tsymbalenko, Ethan V. Munson (2001), "Using HTML Metadata to Find Relevant Imageson the World Wide Web" sponsored by the U. S. Department of Defense. | [7]"HTML Tidy for Python" | [8] Ilia Alshanetsky, "Managing PHP Performance" | [9]JussiMyllymaki, Jared Jackson (2002), "Robust Web Data Extraction with XML Path Expressions" | [10]Ilia Alshanetsky, "PHP &Performance" | [11]"CleanUp your Web pages with HTML tidy", Tidy, W3C | [12]Tidy: (http://tidy.sourceforge.net/) | [13]PECL Tidy: (http://pecl.php.net/tidy) | [14]PHP Tidy: (http://php.net/tidy)