

Tidy

A guide to validating user submitted HTML content



Who am I?

- Michael Tougeron (Mike or Touge)
 - <http://www.grepmymind.com/>
 - Engineering Mgr & Architect for GameSpot.com @ CBS Interactive
<http://www.gamespot.com/>
 - Organizer of the SF PHP & MySQL Meetups

What am I talking about & why?

- What? Tidy. (wow, that was easy)
- Why?
 - Parsing user inputted HTML is a PITA.
 - The user experience matters!
 - XSS sucks.

What is Tidy?

- Tidy is a HTML parser & beautifier – NOT a validator
- HTML Tidy Library is a project that took Dave Ragget's HTML Tidy program and turned it into a set of libraries
 - mod_tidy for Apache
 - Available for C++/Perl/PHP/.NET/more
- PECL extension by John Coggeshall provides PHP API hooks into Tidy

Why validate?

- You should already know why. For those of you who don't:
 - XSS/Security
 - Badly formed HTML breaking your pages
 - Provide a better user experience

Why not bbcode?

- BBcode is not validation.
- Someday, someone, aka your boss, is going to ask you why they can't do JavaScript or embed the latest flash craze.
- If you're not careful, bbcode can still be exploited.

Common methods of validation

- Regular expressions
 - Insanely complicated & doesn't always catch everything.
 - `<(\w+((\s+\S+(\s*=\s*(?:\".*?\"|'.*?'|[\^\\">\s]+))?)+\s*|\s*))>`
- `htmlentities()` or PHP's Filter
 - We wanted to allow HTML, right?
- `strip_tags()`
 - Doesn't check attributes. And IE applies CSS expressions on *closing* tags. (lame)

Common methods of validation

- XML parser
 - Requires strict xhtml by the users & doesn't allow entities. Malformed breaks parser.
- HTML Purifier
 - Good product for small installs. But too s...l...o...w... for high traffic.
- Roll your own (non-Tidy)
 - Most home-rolled code uses a combination of the above.

Common Exploits/Problems

- `<p bad="></p>" onclick="xss">text</p>`
 - strip_tags, xml parser & some regex
- `text</b style="xss:expression">`
 - strip_tags & some regex
- `<p><u>text</p>`
 - xml parser fails with non-specific msg
- `` vs. ``
 - xml parser fails with non-specific msg

Why Tidy?

- Not a validator. It's a parser. This means you're rolling your own rules.
- Tidy can be:
 - Fast
 - Accurate
 - Flexible
 - Customizable
- But it's your code, you can still mess it up.

How the heck do I do this?

- Setup your Tidy configuration
- Decide on a rule format
- Write your tag validator
- Write your attribute validator
- TEST TEST TEST

Seriously... Test your work.

- PHP Unit or other unit test
- QA Team (if you're lucky enough to have one)
- <http://ha.ckers.org/xss.html>
- Find at least one white-hat (if you're lucky enough to know one you can trust)

Configuration Options

- `output-xhtml => true`
 - it's time we all started using it.
- `fix-uri => true`
 - Nice to have non-exploitable URIs
- `word-2000 => true`
 - Lifesaver. I hate MS Word's HTML format
- `show-body-only => true`
 - Not building a page so we only need body

Configuration Options

- drop-proprietary-attributes => true
 - If they've entered something non-standard then they are not playing nice
- And a whole lot more...
 - <http://tidy.sourceforge.net/docs/quickref.html>

Rule format

```
$tags = array(  
  'b' => array('attribs' => array(),  
              'settings' => array(  
                'require_close' => true)  
            ),  
)
```

Rule format

```
'a' => array('attribs' => array(  
    'href' => array('type'=>'url'),  
    'title' => array('type'=>'string'),  
    'onclick'=>array(  
        'type'=>'unrestricted',  
        'auth_level'=>'admin')  
    )  
);
```

Some code (sort of)

```
// Let's initialize tidy with our HTML and config options.
$tidy = tidy_parse_string($html, $config);
// Not let's clean & repair to try and "fix" user errors
$tidy->cleanRepair();
// loop through the cleaned tidy object & validate each tag
foreach ( $tidy->body()->child as $key => $child ) {
    $error = validateTidyNode($child, $new_html);
    if ( $error ) { // merge into larger array of errors }
}
// validateTidyNode is recursive for each child of the child.
// and calls validateHTMLTagAttribute for each attribute

return array('errors'=>$errors, 'html' => $new_html);
```

Tag & Attribute validators

- Doesn't fit in a slide ☹
 - <http://www.grepmy mind.com/2009/01/15/validating-html-with-tidy/>

Credits, Links & other shameless plugs

- Tidy: <http://tidy.sourceforge.net/>
- PECL Tidy: <http://pecl.php.net/tidy>
- PHP Tidy: <http://php.net/tidy>
- HTML Purifier: <http://htmlpurifier.org/>
 - Feature comparison: <http://htmlpurifier.org/comparison.html>
- Michael Tougeron
 - <http://www.grepmymind.com/>
 - michael.tougeron@cbs.com
 - <http://twitter.com/mtougeron>
- SF PHP Meetup: <http://www.meetup.com/sf-php/>
- XSS Cheat sheet: <http://ha.ckers.org/xss.html>