

Question about CGI.pm

Source: <http://coding.derkeiler.com/Archive/Perl/perl.beginners/2008-03/msg00502.html>

- *From:* lesley.binks@xxxxxxxxxxxxxxxx (LesleyB)
 - *Date:* Tue, 25 Mar 2008 07:30:14 -0700 (PDT)
-

Hi

I have been exploring CGI.pm and am of course interested in the HTML escaping procedure.

perldoc CGI throws up this

"By default, all HTML that is emitted by the form-generating functions is passed through a function called `escapeHTML()`:

```
$escaped_string = escapeHTML("unescaped string");  
Escape HTML formatting characters in a string.
```

Provided that you have specified a character set of ISO-8859-1 (the default), the standard HTML escaping rules will be used. The "<" character becomes "<", ">" becomes ">", "&" becomes "&", and the quote character becomes """. In addition, the hexadecimal 0x8b and 0x9b characters, which some browsers incorrectly interpret as the left and right angle-bracket characters, are replaced by their numeric character entities ("‹" and "›"). If you manually change the charset, either by calling the `charset()` method explicitly or by passing a `-charset` argument to `header()`, then all characters will be replaced by their numeric entities, since CGI.pm has no lookup table for all the possible encodings.

The automatic escaping does not apply to other shortcuts, such as `h1()`. You should call `escapeHTML()` yourself on untrusted data in order to protect your pages against nasty tricks that people may enter into guestbooks, etc..

To change the character set, use `charset()`. To turn autoescaping off completely, use `autoEscape(0):`"

and I need to ask some questions about it.

I'm using the OO form in case that makes any difference. Assuming a `'my $qry = new CGI;'`

Question about CGI.pm

The first sentence

"By default, all HTML that is emitted by the form-generating functions is passed through a function called `escapeHTML()`:"

I'm slightly confused by the term 'form-generating' ... does this specifically mean functions such as `start_form`, `checkbox_group`, `submit` and `end_form` and to the exclusion of functions such as `$qry->p(...)` ? Or does it include everything uttered between `$qry->start_form` and `$qry->end_form` which might include a `$qry->div()` or `$qry->p()` ?

The statement later in "The automatic escaping does not apply to other shortcuts, such as `h1()`. You should call `escapeHTML()` yourself on untrusted data in order to protect your pages against nasty tricks that people may enter into guestbooks, etc.." seems to indicate that escaping does not happen and I am tempted to consider "form-generating functions" as those that generate form elements such as radio boxes, pop-up lists, submit buttons and so on.

I think I have understood that if I change my default language to UTF-8 then something like "<" will be translated into a numeric code rather than `<`; But that this will only occur in form-generating functions. Odd how I find just writing the problem out sometimes clarifies things.

I'm mostly working in ISO-8859-1 but would like to 'upgrade' to UTF-8. I have the routine

```
$rslt =~ s/([^\w\s])/sprintf ("&#%d;", ord ($1))/ge;
```

to escape output before committing it to the web page. Any enlightenment as to how to ensure the `ord` function works in a charset dependent way would be gratefully received.

Regards

L.

.