NAME

WWW − World Wide Web Package

SYNOPSIS

```
extract_description( FILE )
extract_meta( FILE, NAME )
hyperlink( LIST )
```

DESCRIPTION

This package provides a utility functions for the World Wide Web to extract descriptions of or meta information from files, and hyperlink text.

SUBROUTINES

The following Perl subroutines are defined and available:

`extract_description( FILE )`

Extracts a description from an HTML or plain text file given by the *FILE* name; *FILE* should be an absolute path. The first `$description::chars` (default: 2048) characters are read. If the file ends in one of the extensions `htm`, `html`, or `shtml`, it is presumed to be an HTML file; if the file ends in `txt`, it is presumed to be a plain text file. Other extensions are not recognized and no description is returned for them.

For HTML files, first, if a `<META NAME="description" CONTENT="...">` or a `<META NAME="DC.description" CONTENT="...">` (Dublin Core) element is found, then the words specified as the value of the `CONTENT` attribute is returned as the description.

Otherwise, all HTML comments, text between `<SCRIPT>`, `<STYLE>`, and `<TITLE>` tags, and all other HTML tags are stripped. If `<AREA ... ALT="...">` or `<IMG ... ALT="...">` elements are found, then the words specified as the value of the `ALT` attributes are extracted.

Finally, for either HTML or plain text files, at most `$description::words` (default: 50) are returned.

`extract_meta( FILE, NAME )`

Extracts the value of the `CONTENT` attribute from a `META` element having the given `NAME` attribute from an HTML file given by the *FILE* name; *FILE* should be an absolute path. The file must end in one of the extensions `htm`, `html`, or `shtml` to be considered an HTML file. The first `$description::chars` (default: 2048) characters are read. The characters are cached between consecutive calls using the same filename.

`hyperlink( LIST )`

Adds hyperlinks to strings: that is strings that contain substrings that are valid URLs (according to RFC 1630) have the appropriate HTML tags ‘‘wrapped’’ around them so that they will be selectable when displayed in a browser. The `ftp`, `gopher`, `http`, `https`, `mailto`, `news`, `telnet`, and `wais` URLs are recognized. Example:

```
Read all about it at
http://www.usatoday.com/
```

becomes:

```
Read all about it at
<A HREF="http://www.usatoday.com/">http://www.usatoday.com/</A>
```

SEE ALSO

perl(1)

Tim Berners-Lee. ‘‘Universal Resource Identifiers in WWW,’’ *Request for Comments 1630*, Network Working Group of the Internet Engineering Task Force, June 1994.

Tim Berners-Lee, Larry Masinter, and Mark McCahill. ‘‘Uniform Resource Locators (URL),’’ *Request for*

Comments 1738, Network Working Group, 1994.

Dave Raggett, Arnaud Le Hors, and Ian Jacobs. ‘‘Notes on helping search engines index your Web site,’’ *HTML 4.0 Specification, Appendix B: Performance, Implementation, and Design Notes*, World Wide Web Consortium, April 1998.

−−. ‘‘Objects, Images, and Applets: How to specify alternate text,’’ *HTML 4.0 Specification, §13.8*, World Wide Web Consortium, April 1998.

Dublin Core Directorate. ‘‘The Dublin Core: A Simple Content Description Model for Electronic Resources.’’

Larry Wall, et al. *Programming Perl, 3rd ed.*, O’Reilly & Associates, Inc., Sebastopol, CA, 2000.

AUTHOR

Paul J. Lucas <*pauljlucas@mac.com*>