Table of Contents

Appendix B: Encode URLs With UTF8

Use internationalised characters within WikiWords and attachment names

Current Status

To simplify use of internationalised characters within WikiWords and attachment names, Foswiki supports UTF-8 URLs, converting on-the-fly to virtually any character set, including ISO-8859-*, KOI8-R, EUC-JP, and so on.

Support for UTF-8 URL encoding avoids having to configure the browser to turn off this encoding in URLs (the default in Internet Explorer, Opera Browser and some Mozilla Browser URLs) and enables support of browsers where only this mode is supported (e.g. Opera Browser for Symbian smartphones). A non-UTF-8 site character set (e.g. ISO-8859-*) is still used within Foswiki, and in fact pages are stored and viewed entirely in the site character set - the browser dynamically converts URLs from the site character set into UTF-8, and Foswiki converts them back again.

System requirements are updated as follows:

- ASCII or ISO-8859-1-only sites do not require any additional CPAN modules to be installed.
- Perl 5.8 sites using any character set do not require additional modules, since CPAN:Encode is installed as part of Perl.
- This feature still works on Perl 5.005_03 as per SystemRequirements, or Perl 5.6, as long as CPAN:Unicode::MapUTF8 is installed.

The following 'non-ASCII-safe' character encodings are now excluded from use as the site character set, since they interfere with Foswiki markup: ISO-2022-*, HZ-*, Shift-JIS, MS-Kanji, GB2312, GBK, GB18030, Johab and UHC. However, many multi-byte character sets work fine, e.g. EUC-JP, EUC-KR, EUC-TW, and EUC-CN. In addition, UTF-8 can already be used, with some limitations, for East Asian languages where EUC character encodings are not acceptable.

It's now possible to override the site character set defined in the `{SiteLocale}` setting in configure - this enables you to have a slightly different spelling of the character set in the server locale (e.g. 'eucjp') and the HTTP header sent to the browser (e.g. 'euc-jp').

Details of Implementation

URLs are not allowed to contain non-ASCII (8th bit set) characters:
http://www.w3.org/TR/html4/appendix/notes.html#non-ascii-chars

UTF-8 URL translation to virtually any character set is supported, but full UTF-8 support (e.g. pages in UTF-8) is not supported yet.

The code automatically detects whether a URL is UTF-8 or not, taking care to avoid over-long and illegal UTF-8 encodings that could introduce security issues (tested against a comprehensive UTF-8 test file, which IE 5.5 fails quite dangerously, and Opera Browser passes). Any non-ASCII URLs that are *not* valid UTF-8 are then assumed to be directly URL-encoded as a single-byte or multi-byte character set (as now), e.g. EUC-JP.

The main point is that you can use Foswiki with international characters in WikiWords without changing your browser setup from the default, and you can also still use Foswiki using non-UTF-8 URLs. This works on any Perl version from 5.005_03 onwards. You can have different users using different URL formats transparently on the same server.

UTF-8 URLs are automatically converted to the current {Site}{Charset}, using modules such as if needed.

Foswiki generates the whole page in the site charset, e.g. ISO-8859-1 or EUC-JP, but the browser dynamically UTF-8 encodes the attachment's URL when it's used. Since Apache serves attachment downloads without Foswiki being involved, Foswiki's code can't do its UTF-8 decoding trick, so Foswiki URL-encodes such URLs in ISO-8859-1 or whatever when generating the page, to bypass this URL encoding, ensuring that the URLs and filenames seen by Apache remain in the site charset.

Related Topics: AdminDocumentationCategory
Edit | Attach | Print version | History: %REVISIONS% | Backlinks | Raw View | More topic actions
Topic revision: r1 - 12 Sep 2009 - 04:09:53 - ProjectContributor