Table of Contents

Appendix B: Encode URLs With UTF8

Use internationalised characters within WikiWords and attachment names

This topic addresses implemented UTF-8 support for URLs only. The overall plan for UTF-8 support for TWiki is described in TWiki:Codev.ProposedUTF8SupportForI18N.

Current Status

To simplify use of internationalised characters within WikiWords and attachment names, TWiki now 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 TWiki, 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 TWiki 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 TWikiSystemRequirements, 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 TWiki 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 - see TWiki:Codev.ProposedUTF8SupportForI18N.

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').

This feature should also support use of Mozilla Browser with TWiki:Codev.TWikiOnMainframe (as long as mainframe web server can convert or pass through UTF-8 URLs) - however, this specific combination is not tested. Other browser-server combinations should not have any problems.

Please note that use of UTF-8 as the site character set is not yet supported - see Phase 2 of TWiki:Codev.ProposedUTF8SupportForI18N for plans and work to date in this area.

This feature is complete in TWiki releases newer than February 2004.

Note for skin developers: is no longer required (TWiki:Plugins.InternationalisingYourSkin).

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

The overall plan for UTF-8 support for TWiki is described in two phases in TWiki:/Codev.ProposedUTF8SupportForI18N - this page addresses the first phase, in which UTF-8 is supported for URLs only.

UTF-8 URL translation to virtually any character set is supported as of TWiki Release 01 Sep 2004, but full UTF-8 support (e.g. pages in UTF-8) is not supported yet - this will be phase 2.

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 TWiki:Codev.MajorSecurityProblemWithIncludeFileProcessing (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 TWiki with international characters in WikiWords without changing your browser setup from the default, and you can also still use TWiki using non-UTF-8 URLs. This works on any Perl version from 5.005_03 onwards and corresponds to Phase 1 of TWiki:Codev.ProposedUTF8SupportForI18N. 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 CPAN:Encode if needed.

TWiki 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 TWiki being involved, TWiki's code can't do its UTF-8 decoding trick, so TWiki 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.

TWiki:Codev.TWikiOnMainframe uses EBCDIC web servers that typically translate their output to ASCII, UTF-8 or ISO-8859-1 (and URLs in the other direction) since there are so few EBCDIC web browsers. Such web servers don't work with even ISO-8859-1 URLs if they are URL encoded, since the automated translation is bypassed for URL-encoded characters. For TWiki on Mainframe, TWiki assumes that the web server will automatically translate UTF-8 URLs into EBCDIC URLs, as long as URL encoding is turned off in TWiki pages.

Testing and Limitation

It should work with TWiki:Codev.TWikiOnMainframe. Tested with IE 5.5, Opera 7.11 and Mozilla (Firebird 0.7).

Opera Browser on the P800 smartphone is working for page viewing but leads to corrupt page names when editing pages.

For up to date information see TWiki:Codev.EncodeURLsWithUTF8

Related Topics: AdminDocumentationCategory

This topic: TWiki > AppendixEncodeURLsWithUTF8
Topic revision: r4 - 2005-03-27 - TWikiContributor

Note: Please contribute updates to this topic on TWiki.org at TWiki:TWiki.AppendixEncodeURLsWithUTF8