# UTF-8 with Gentoo/Linux

Lars Weiler <pylon@gentoo.org>

Gentoo Foundation

FOSDEM – Free and Open Source Software Developers' European Meeting
Brussels
February 27, 2005

# What it UTF-8?

**UTF-8** (8-bit Unicode Transformation Format) is a lossless, variable-length character encoding for Unicode [...]. It uses groups of bytes to represent the Unicode standard for the alphabets of many of the world's languages. UTF-8 is especially useful for transmission over 8-bit mail systems.

From Wikipedia, the free encyclopedia.
http://en.wikipedia.org/wiki/UTF-8

see also RFC 3629 (UTF-8, a transformation format of ISO 10646)

# Current character encodings

- ISO-8859-1 (latin1) or ISO-8859-15 (latin9) for Western Europe with a maximum of 8 bit characters
- ISO-8859-2 (latin2) for Central and Eastern Europe, ISO... etc. for further character-based encodings
- KOI8-R/U for cyrillic
- ISO-2022-JP in Japan
- ...etc.

# Why using UTF-8?

- More characters!
- No convertion problems
- Pep up text with special characters (e.g. Klingonic signs)
- Writing texts in different languages and character sets
- Internationally UTF-8 will become a leader in character encoding

# How does UTF-8 work?

- Full compatibility to ASCII in the first seven bits
- If the 8th bit is a 1, another byte will be “appended”

```
0xxxxxxx → 127 Characters

110xxxxx 10xxxxxx → 1920 Characters

1110xxxx 10xxxxxx 10xxxxxx → 63488 Characters

11110xxx 10xxxxxx 10xxxxxx 10xxxxxx → 1.048.576 Char.
```

- Theoretically more bytes could be used but UTF-8 is limited to 4 bytes

# Activating UTF-8

- UTF-8 locales must be existant
- During compilation of `glibc` they will be created automatically
- `locale` shows the currently set locales
- `locale -a` shows all available locales

# Example for locale

```
$ locale
LANG=german
LC_CTYPE="de_DE.utf8"
LC_NUMERIC="de_DE.utf8"
LC_TIME="de_DE.utf8"
LC_COLLATE="de_DE.utf8"
LC_MONETARY="de_DE.utf8"
LC_MESSAGES="de_DE.utf8"
LC_PAPER="de_DE.utf8"
LC_NAME="de_DE.utf8"
LC_ADDRESS="de_DE.utf8"
LC_TELEPHONE="de_DE.utf8"
LC_MEASUREMENT="de_DE.utf8"
LC_IDENTIFICATION="de_DE.utf8"
LC_ALL=de_DE.utf8
```

# Example for `locale -a`

```
$ locale -a
C
de_DE
de_DE@euro
de_DE.iso88591
de_DE.iso885915@euro
de_DE.utf8
deutsch
en_US
en_US.iso88591
en_US.utf8
german
POSIX
```

# Compiling glibc with special locales

- Enable the userlocales USE-flag
- List in /etc/locales.build the to be built locales:

```
en_US/ISO-8859-1
en_US.UTF-8/UTF-8
de_DE/ISO-8859-1
de_DE@euro/ISO-8859-15
de_DE.UTF-8/UTF-8
```

# Enabling UTF-8 globally

- Write into `/etc/env.d/02locale` `LANG` and `LC_ALL` (and further Variables):

```
LANG=”german”

LC_ALL=”de_DE.utf8”

GDM_LANG=”de_DE.utf8”
```

- Then

```
# env-update && source /etc/profile
```

- And probably do a re-login

# Better enable UTF-8 on a per user basis

- Add to your `~/.profile` (or `~/.login` for a C shell):

```
export LANG=”de_DE.utf-8”
export LC_ALL=”de_DE.utf-8”
```

- `re-login`

# USE-flag unicode

- Some packages need to be compiled with the USE-flag unicode enabled
- Just add unicode to the USE-Variable in /etc/make.conf

# Additional console configuration

- Enable in `/etc/rc.conf`
  `UNICODE="yes"`
- Choose a UTF-8-compatible `CONSOLEFONT` (if there is one)
- Prepend `KEYMAP` with `-u`:
  `KEYMAP="-u de"`
- If there are problems with non-ASCII characters, run `unicode_start`

# X-Terminal with UTF-8

- KDE's Konsole and gnome-terminal provide UTF-8 in the settings
- xterm is UTF-8 capable, but needs some settings in `~/.Xresources` or the correct command line parameters
- Fast and lightwight UTF-8 Terminal: `rxvt-unicode (urxvt)`

# Special settings: less

- less somehow does not use the set locales
- Edit in /etc/env.d/70less

```
LESSCHARSET="utf-8"
```

- Don't forget
  env-update && source /etc/profile

# Further settings

- Most modern applications use the locales and change to the appropriate character set
- Problem areas:
  - gtk1
  - bash, readline (Upgrade to bash-3 and readline-5)
  - Fonts without Unicode-characters
  - Applications without character-rewrite

# Files with special characters

- Activate in the kernel `CONFIG_NLS_UTF8`
- `CONFIG_NLS_DEFAULT` should be set to `utf-8`
- With app-text/convmv filenames could be converted to UTF-8
- Samba-3 talks UTF-8

# Vim as UTF-8 Editor

- Evaluate the settings:
  - `:set encoding=utf-8`
  - `:set fileencoding=utf-8`
- Vim automatically converts files to UTF-8 — the important setting is `fileencoding`
- Enter the UTF-8 characternumber with ctrl-v-u <code>
  - `ctrl-v-u 03c0` → π

# UTF-8 with the example of Gentoo Documentation

- No nasty convertions
- Escape-sequences or entities are not needed (e.g. &#03c3;)
- Easier writing of text
- Editors could change the layout even if they don't speak the language used for the document

# “deine umlaute sind kaputt!!1!elf1eins!”

- Currently the biggest problem is IRC
- UTF-8 isn't accepted by old IRC-stagers
- Workaround:
  - Write out Umlauts (ä -> ae)
  - Set IRC-Client to latin1/latin9
  - Use /recode
  - Wait until UTF-8 will be accepted more widely…

# E-Mail with UTF-8

- After the usual beginner problems, nowadays quite every Mail User Agents supports UTF-8
- Internally UTF-8 messages are transformed to the locale set by the user
- Korean SPAM could now be displayed correctly! (ann.: benefit?)

# Summary

- UTF-8 is on the way becoming standard in Gentoo (estimated in 2005) — other Linux Distributions already switched to UTF-8
- Many applications don't cause troubles
- Here and there some configuration is needed
- The changeover to UTF-8 isn't (no longer) that hard

# Resources

- Gentoo Linux Documentation: Using UTF-8 with Gentoo:
  http://www.gentoo.org/doc/en/utf-8.xml
- UTF-8 Sampler:
  http://www.columbia.edu/kermit/utf8.html
- Markus Kuhn: The UTF-8 and Unicode FAQ for Unix/Linux:
  http://www.cl.cam.ac.uk/~mgk25/unicode.html
- Project UTF-8, freedesktop.org:
  http://freedesktop.org/Software/utf-8
- 1½ years development for the integration of UTF-8:
  http://bugs.gentoo.org/show_bug.cgi?id=18375

# Finish

Thanks you for your attention!