



# UTF-8 mit Gentoo/Linux

Lars Weiler <[pylon@gentoo.org](mailto:pylon@gentoo.org)>

Gentoo Foundation und  
Förderverein Gentoo e.V.

Gentoo User Meeting Ruhrpott  
Oberhausen  
5. November 2004



# Was ist UTF-8?

**UTF-8** (Abk. für 8-bit Unicode Transformation Format) ist die fortschrittlichste und populärste Kodierung für Unicode-Zeichen; dabei wird jedem Unicode-Zeichen eine speziell kodierte Bytekette von variabler Länge zugeordnet. Es unterstützt bis zu 4 Byte auf die sich wie bei allen UTF Formaten alle 1.114.112 Unicode Zeichen abbilden lassen.

aus Wikipedia, der freien Enzyklopädie

<http://de.wikipedia.org/wiki/UTF-8>

siehe auch RFC 3629: UTF-8, a transformation format of ISO 10646



# Zeichensysteme heute

- ISO-8859-1 (latin1) bzw. ISO-8859-15 (latin9) für Mitteleuropa mit 8 Bit maximalen Zeichen
- ISO-8859-2 (latin2) für Osteuropa, ISO... usw. für andere Zeichen
- KOI8-R/U für kyrillische Zeichen
- ISO-2022-JP in Japan
- ...usw. Siehe auch

<http://de.wikipedia.org/wiki/Kategorie:Zeichenkodierung>



# Warum UTF-8 verwenden?

- Mehr Zeichen!
- Keine Konvertierungsprobleme
- Text mit besonderen Zeichen aufpeppen (bspw. Klingonisch)
- Texte in verschiedenen Sprachen und Zeichensystemen verfassen
- International wird sich Unicode mit UTF-8 durchsetzen



# Wie funktioniert UTF-8?

- Volle Kompatibilität zu ASCII in den ersten 7 Bit
- Ist das 8. Bit eine 1 wird ein weiteres Byte “angehängen”

0xxxxxxx → 127 Characters

110xxxxx 10xxxxxx → 1920 Characters

1110xxxx 10xxxxxx 10xxxxxx → 63488 Characters

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

- Mehr ist möglich, jedoch wurde UTF-8 auf 4 Byte beschränkt



# UTF-8 aktivieren

- UTF-8 locales müssen existieren
- Beim compilen der glibc werden diese inzwischen automatisch erzeugt
- `locale` zeigt die aktuell eingestellten locales an
- `locale -a` zeigt alle zur Verfügung stehenden locales an



# Beispiel zu 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
```



# Beispiel zu 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
```



# glibc mit speziellen locales compilieren

- Das USE-flag `userlocales` setzen
- In `/etc/locales.build` die gewünschten locales eintragen:

`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`



# UTF-8 global aktivieren

- In `/etc/env.d/02locale` LANG und LC\_ALL (und weitere Variablen) entsprechend setzen:

```
LANG="german"
```

```
LC_ALL="de_DE.utf8"
```

```
GDM_LANG="de_DE.utf8"
```

- Danach

```
# env-update && source /etc/profile
```
- Und gegebenenfalls neu einloggen



# USE-flag unicode

- Manche Pakete müssen mit dem USE-flag `unicode` kompiliert werden
- Einfach in der `/etc/make.conf` zur USE-Variable hinzufügen



# Zusätzliche Konfiguration für die Console

- In `/etc/rc.conf` `UNICODE="yes"` einschalten
- Eine UTF-8-kompatible `CONSOLEFONT` wählen (Anm.: gibt es schon eine?)
- Als Workaround `CONSOLETRANSLATION="8859-1_to_uni"` aktivieren
- Bei Problemen mit der Anzeige `unicode_start` ausführen



# X-Terminal mit UTF-8

- KDEs Konsole und gnome-terminal bieten UTF-8 unter den Einstellungen
- xterm ist grundsätzlich UTF-8 fähig, benötigt aber einige Einstellungen in `~/.Xresources`
- Schnelles, leichtes UTF-8-fähiges Terminal: `rxvt-unicode` (`urxvt`)



# Spezialeinstellung: less

- less liest aus irgendeinem Grund nicht die eingestellten locales aus
- Hier muss in `/etc/env.d/70less` manuell nachgeholfen werden:

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

- `env-update && source /etc/profile`  
nicht vergessen!



# Sonstige Einstellungen

- Viele moderne Applikationen lesen die locales aus und stellen den zu verwendenden Zeichensatz darauf ein
- Problemzonen:
  - gtk1
  - bash, readline (Upgrade auf bash-3 respektive readline-5)
  - Schriften ohne Unicode-Unterstützung
  - Programme ohne character-rewrite



# Dateien mit Sonderzeichen

- Im Kernel muss `CONFIG_NLS_UTF8` aktiviert sein
- `CONFIG_NLS-DEFAULT` sollte auf `utf-8` stehen
- Mit `app-text/convmv` lassen sich Dateinamen in UTF-8 umwandeln
- Samba-3 spricht UTF-8



# Vim als UTF-8 Editor

- Einstellungen prüfen:
  - `:set encoding=utf-8`
  - `:set fileencoding=utf-8`
- Vim wandelt automagisch Dateien intern in UTF-8 um — zum Speichern ist `fileencoding` ausschlaggebend
- Direktes Eingeben des UTF-8-Zeichencodes mittels `ctrl-v-u <code>`
  - `ctrl-v-u 03c0` → π



# UTF-8 am Beispiel der Gentoo-Dokumentation

- Keine lästigen Umwandlungen
- Escape-Sequenzen bzw. Entities sind nicht notwendig (z.B. `&#03c3;`)
- Einfacheres Schreiben von Fließtext
- Der Sprache nicht mächtige Editoren können dennoch Änderungen am Layout vornehmen



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

- Größte Problemzone ist immer noch das IRC
- UTF-8 wird bei alteingesessenen IRC-Hasen nicht gerne gesehen
- Abhilfe:
  - Umlaute ausschreiben
  - IRC-Client auf latin1/latin9 einstellen
  - Automatisch recoden lassen
  - Warten bis UTF-8 zum Standard wird...



# E-Mail mit UTF-8

- Nach anfänglichen Problemen unterstützt nun nahezu jeder Mail User Agent UTF-8
- Intern werden UTF-8 Mails in die vom Benutzer eingestellte locale zur Anzeige umgewandelt
- Koreanischer SPAM wird sauber angezeigt! (Anm.: Nutzen?)



# Zusammenfassung

- UTF-8 ist auf dem Weg zum Standard bei Gentoo (vorraussichtlich noch in 2005) — andere Linux-Distributionen haben bereits diesen Schritt vollzogen
- Viele Applikationen bereiten keine großen Probleme
- An manchen Stellen ist ein wenig Konfiguration notwendig
- Der Umstieg ist gar nicht (mehr) so schwer



# Weiterführende Dokumentation

- 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>
- HOWTO: Using UTF-8 on Gentoo (etwas veraltet):  
<http://forums.gentoo.org/viewtopic.php?t=166984>
- Project UTF-8, freedesktop.org:  
<http://freedesktop.org/Software/utf-8>
- 1,5 Jahre Entwicklungszeit bis zur Integration in Gentoo:  
[http://bugs.gentoo.org/show\\_bug.cgi?id=18375](http://bugs.gentoo.org/show_bug.cgi?id=18375)



# Ende

Vielen Dank für die Aufmerksamkeit und viel  
Spaß bei der Umstellung auf UTF-8!

...aber vorher: Ran an die Schnitzelplatte!!!