# Unicode & IDN

- Hintergründe
- Wie funktioniert Unicode?
- Im Alltag
  - typische Probleme
  - Webseite mit IDN
- Links

Thomas Kühne <thomas@kühne.name>

# Sprache & Schrift

| | |
|---|---|
| Deutsch: | Latein |
| Französisch: | Latein |
| Japanisch: | Katakana<br>Hiragana<br>Kangxi |
| Insgesamt: | 93+ |

# „Zeichensätze”

**Europa**

- z.B. ISO-8859-1
- 1 Zeichen: 1 Byte
- $2^8 = 256$ Zeichen
- 11 Zeichensätze typisch für EU

**Asien**

- z.B. ISO-2022-JP
- 1 Zeichen: 1-2 Byte
- Escape-Sequenzen

# Locale: Schauplatz

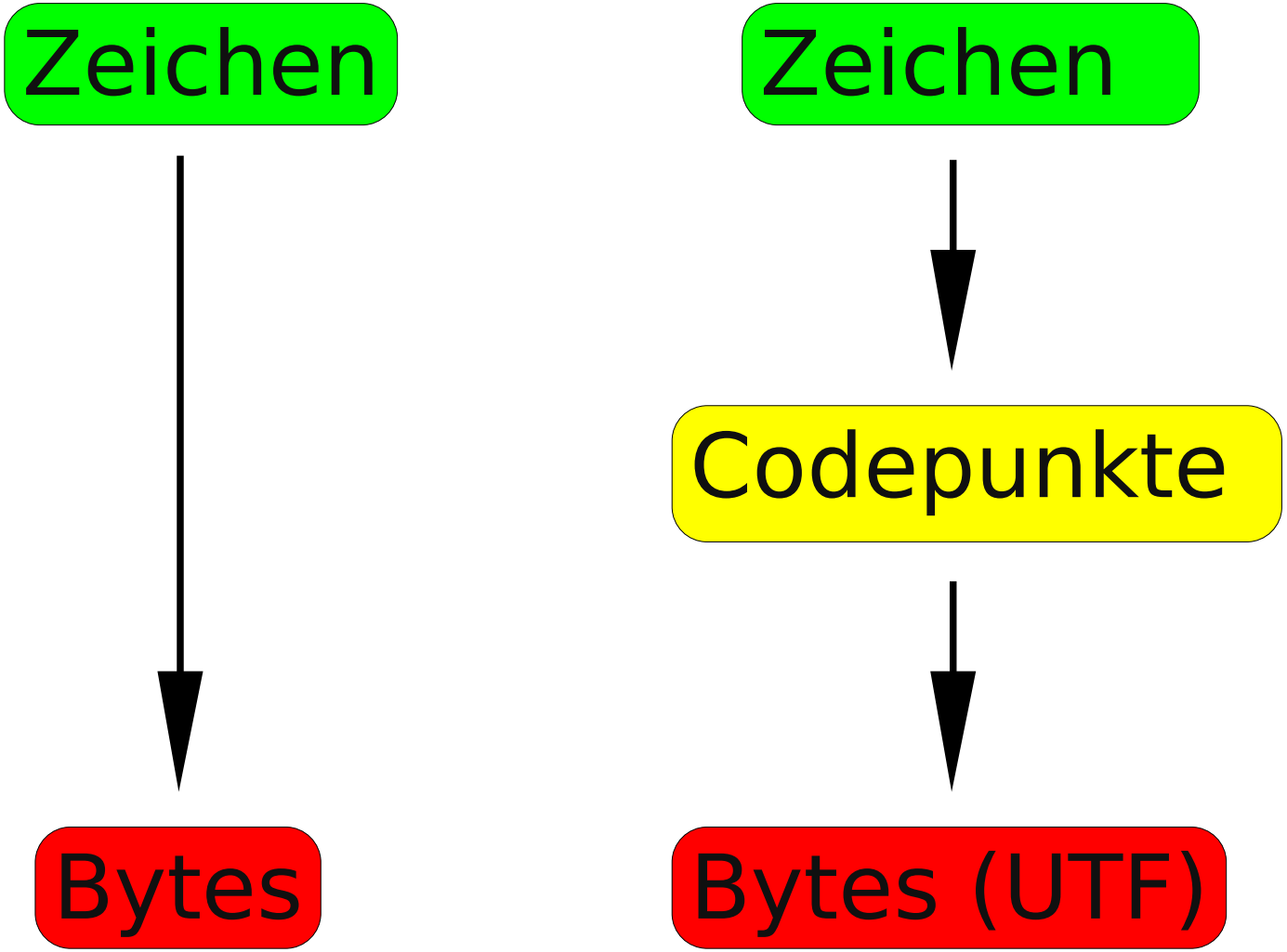
```
> LC_ALL=de_DE.UTF8 date +%c
So 14 Mär 2010 10:17:55 CET

> LC_ALL=de_AT.UTF8 date +%c
Son 14 Mär 2010 10:17:55 CET

> LC_ALL=zh_TW.UTF8 date +%c
西元 2010 年 03 月 14 日 （週日）10 時 17 分 55 秒
```

```
> man locale
```

# ISO 10646 / Unicode

- **ISO/IEC 10646** seit 1989
  - „Universal Character Set“ UCS
  - Mitglieder: staatliche Organisationen
    z.B.: AFNOR(fr), ANSI(us), BSI(gb), DIN(de)
- **Unicode Consortium** seit 1987
  - 1991 „The Unicode Standard 1.0“
  - UCS + BIDI, Sortierung, etc.
  - Mitglieder: vorrangig private Organisationen
    z.B.: Adobe, DENIC, HP, IBM, Yahoo!

# Genereller Aufbau



# Unicode Transformation Format (UTF)

- Byte Order Mark (BOM)
  Codepunkt: U+FEFF
- Big Endian (BE) - Little Endian (LE)
- Kodierungen:
  UTF-8
  UTF-16 BE  UTF-16 LE
  UTF-32 BE  UTF-32 LE

# UTF-8

U+000000 – U+00007F

→ 0xxxxxxx

U+000080 – U+0007FF

→ 110yyyxx 10xxxxxx

U+000800 – U+00FFFF

→ 1110yyyy 10yyyyxx 10xxxxxx

U+010000 – U+10FFFF

→ 11110zzz 10zzyyyy 10yyyyxx 10xxxxxx

# UTF-16

U+000000 – U+00FFFF

→ yyyyyyyy xxxxxxxx

U+010000 – U+10FFFF

→ 110110zz zzyyyyyy High-Surrogate

110111yy xxxxxxxx Low-Surrogate

# typischer UTF Einsatz

- *nix
  - UTF-8 ohne BOM
  - wchar_t: UTF-32 ohne BOM
- Windows
  - UTF-16 mit BOM
  - wchar_t: UTF-16 ohne BOM
- Java
  - UTF-16 ohne BOM
  - <1.5: UTF-8 mit UTF-16 Surrogates!

# Normalization Forms

- **NFC** - vorschriftsmäßige Zerlegung gefolgt von vorschriftsmäßiger Zusammensetzung
  ä U+00E4
- **NFD** - vorschriftsmäßig Zerlegung
  a+¨ U+0061 U+0308
- **NFKD** - kompatible Zerlegung
- **NFKC** - kompatible Zerlegung gefolgt von vorschriftsmäßiger Zusammensetzung

# IDN

beißend.Küh-ne.name

Nameprep

beissend.küh-ne.name

Punycode

beissend.xn--kh-ne-kva.name

# Webserver

http://kühne.name/bücher

IDN / escaped UTF-8

http://xn--khne-0ra.name/b%C3%BCcher

# Webseite

- HTTP-Header
- (XML-Header)
- HTML-Header
- Eigentlicher Inhalt

  ü  ü  ü  ü
- Locale
- Browsereinstellungen

# Falsche Kodierung

- UTF-8 als ISO-8895-1: GrÃ¼ÃŸe
- UTF-8 als UTF-16: 片벁?
- UTF-16 als UTF-8: ??G□r□?□?□e□
- UTF-16 als ISO-8859-1: ÿþG□r□ü□ß□e□
- ISO-8859-1 als UTF-8 Gr??e
- ISO-8859-1 als UTF-16 賢?

# Schrift-Zeichensatz-Hack

| | | | | |
|---|---|---|---|---|
| ě | . | / | é | ā |
| á | ǎ | à | ō | ó |
| ǒ | ò | ē | : | ; |
| ( | è | ) | ? | í |
| A | B | C | D | E |

| | | | | |
|---|---|---|---|---|
| - | . | / | 0 | 1 |
| 2 | 3 | 4 | 5 | 6 |
| 7 | 8 | 9 | : | ; |
| < | = | > | ? | @ |
| A | B | C | D | E |

# Bidi (bidirektionaler Text)

→ Demo

# Knackpunkte

- BOM am Anfang des Streams?
- Ist die Eingabe "normalisiert"?
- Welche Zeilenenden liegen vor?
  *nix: U+000A DOS: U+000D U+000A
  Mac: U+000D AIX: U+0085
  Zeile: U+2028 Paragraph: U+2029
- 1 byte ≠ 1 Codepunkt
- 1 Codepunkt ≠ 1 Zeichen

# Links

- Unicode
  http://www.unicode.org/
- UTF-8
  http://www.ietf.org/rfc/rfc3629.txt
- ICU
  http://www.ibm.com/software/globalization/icu/
- BIDI
  http://fribidi.org/
- IDN
  http://www.gnu.org/software/libidn/

# Fragen?