CIDLeS Summer School 2014

Characters, encodings, fonts and keyboards

Pavel Mihaylov

Character sets and encodings

- Character set: … A, B, C, D …
- Encoding: … A→65, B→66, C→67, D→68 …
- Historically each character set had exactly one encoding

Character set soup heritage

- Every script has one or more separate encodings
- Software has to know the encoding
- Cyrillic encodings:
  - CP866, MIK, CP855, CP1251, KOI8-R, KOI8-U, ISO8859-5, MacCyrillic

ÒÏÓÓÉÑ, íÏÓË×Á, 119415

ÐÒ. ÷ÅÒÎÁÄÓËÏÇÏ, 37,

Ê.1817-1,

ÐÌÅÔÎÅ×ÏÅ ÓÏÇÌÁÛÅÎÉÅ.

RUSSIE.

Mojibake

Zeichensalat

кракозябры

- Mojibake (文字化け); lit. “character transformation”), from the Japanese 文字 (moji) “character” + 化け (bake) “transform”, is the phenomenon of symbol aliasing, and a name for the resulting garbled text, arising from systematic errors along a text encoding-transfer-decoding chain.

乱码

маймуница

krzaczki

òÏÓÓÉÑ, íÏÓË×Á, 119415
ÐÒ. ÷ÅÒÎÁÄÓËÏÇÏ, 37,
Ë.1817-1,
ðÌÅÔÎÅ×ÏÊ ó×ÅÔÌÁÎÅ

```
iconv -t iso8859-1 | iconv -f koi8-r
```

Россия, Москва, 119415
пр. Вернадского, 37,
к.1817-1,
Плетневой Светлане

ÒÏÒÒÉÑ, ÍÏÓË×Á, 119415

Ð/Ò. ÅÒÎÁÄÓÉÇÉ, 37,

Ê.1817-1,

ÐÒÅÄÎÅ×ÏÅ ÓÏÏÂÝÅÎÉÅ.

RUSSIE.

Unicode: the magical solution

- A *single character set* that represents all the scripts
  - In current use
  - Historical
  - Fictional
- *Multiple encodings* to represent Unicode
  - UTF-8, UTF-16, UTF-32, UTF-7, etc.

Fonts

- Is Unicode enough?
- Is a font *supporting* the code points for your script enough?
- One Unicode code point may correspond to multiple *glyphs*
- A sequence of Unicode code points may correspond to a single glyph

Keyboards

- MacOSX:
  - Ukelele
- Windows:
  - Microsoft Keyboard Layout Creator
- Linux:
  - X11 keyboard definition files (plain text)
- Android:
  - Key layout files (plain text) + creating an apk

Is all that enough?

- Not always.
- You need to take into account the social aspect.
  - e.g. in India “computers are in English”