Text Encoding

Introduction to Digital Scholarly Editing

Gustavo Fernández Riva

January 2019

Fellow CONICET (Argentina)
Visiting Scholar IEM – FCSH – Universidade Nova Lisboa (Portugal)

The big question

What is TEXT?

Computational Encoding

For a computer all information must be encoded as binary bits (0-1).
The minimal unit of textual information is a *character*.

ASCII

The American Standard Code for Information Interchange (ASCII) was developed in the 1960's based on telegraphic communication.
It used 7 bit-integers for each characters, which amounts to 128 possible combinations (2^7).
32 characters were command characters (not printable). All uppercase characters came before lowercase.
Examples:

- 100 0001 = A (dec. 65, hex. 41)
- 110 0001 = a (dec. 97, hex. 61)
- 010 0000 = space (dec. 32, hex. 20)
- 000 1010 = line feed (dec. 10, hex A)
- 000 1101 = carriage return (dec. 13, hex. D)

8 bits then allowed to place one more, expanding from 128 to 256 (2^8).
Many (mutually incompatible) Extended Encodings were implemented.

Unicode

Since 1991. Each Character is defined by a **code point**, ie. a number (usually expressed in hexadecimal: U+004A = J). Can encode up to 1,111,998, currently 137,439 are used. Multiple encodings of code points are possible (UTF-8, UTF-16, UTF-32). Most commong is **UTF-8**[1]:

- encodes code points in 1 to 4 bytes.
- backwards compatible with ASCII
- ASCII compatible 1 byte= 0x xx xx xx
- 2 bytes = 11 0x xx xx 10 xx xx xx
- 3 bytes = 11 10 xx xx 10 xx xx xx 10 xx xx xx

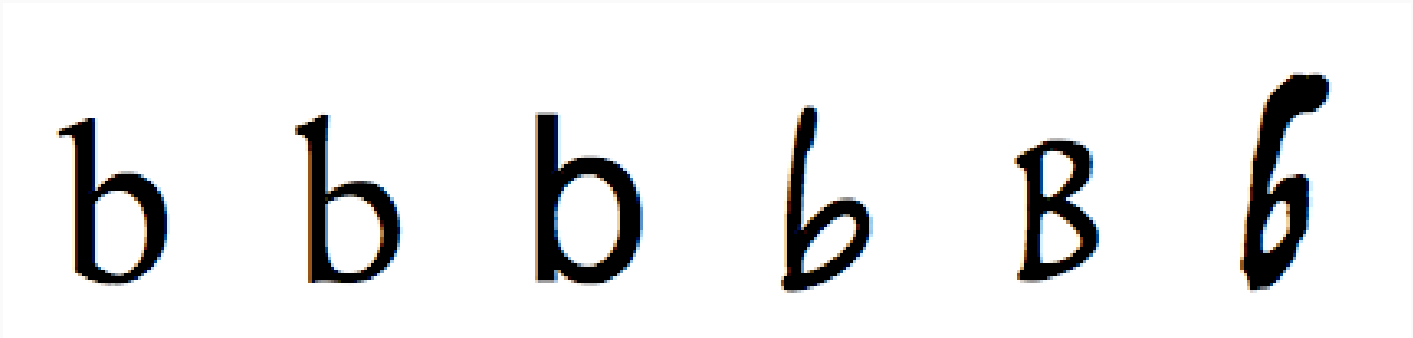
[1] Always use UTF-8!!!!

Font / Encoding

- **Encoding**: Binary representation of characters
- **Font**: Visual representation of characters on screen or print.

Character Latin Capital Letter B:

- Unicode CodePoint: Bin: 1000010, Dec: 66, Hex: 42.
- Unicode UFT-8: 01000010
- Unicode utf-16: 00000000 01000010



Programs and Formats

TEXTUAL DATA ONLY

Text Editors:

- **Basic**: Microsoft Notepad, Apple TextEdit, etc.
- **Advanced** (include XML, Python, etc.): Sublime, BBEdit, Emacs, Atom, etc.

Plain Text:

- .txt (.xml, .html, .py)

TEXT + OTHER

Text Processors:

- Microsoft Office - Word
- LibreOffice - Writer

Not plain Text:

- .doc, .docx, .epub, .pdf, .odt,.jpeg

Exercise

Transcribe into a plain text format:

- https://archive.org/details/whipperginny00gravuoft/page/1
- https://archive.org/details/camoes00alme/page/n11
- https://archive.org/details/camoes00alme/page/n17
- https://archive.org/details/bub_gb_h1qUMLaMp3wC/page/n11