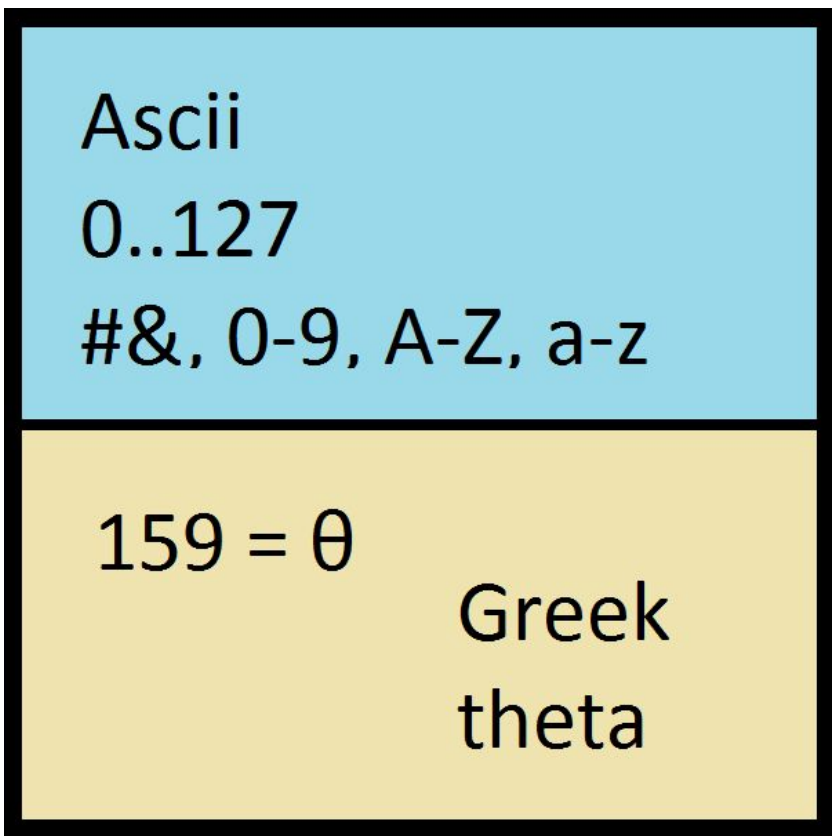
Unicode – What is it?

Juraj Pivovarov, M.Sc.
Metadata Innovations Inc.

Byte Plane
8 bits = 256 combinations

- Everyone agrees on Ascii
- No one agrees on High Ascii

CONFLICT: Soon, it's all Greek!

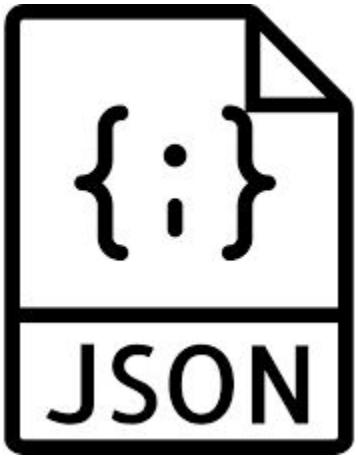


What is Unicode?

- Universal Coded Character Set
- In theory: 1,114,112 code points possible!
- In practice: 136,755 code points

Encodings

Relational Database



Encoding Unicode

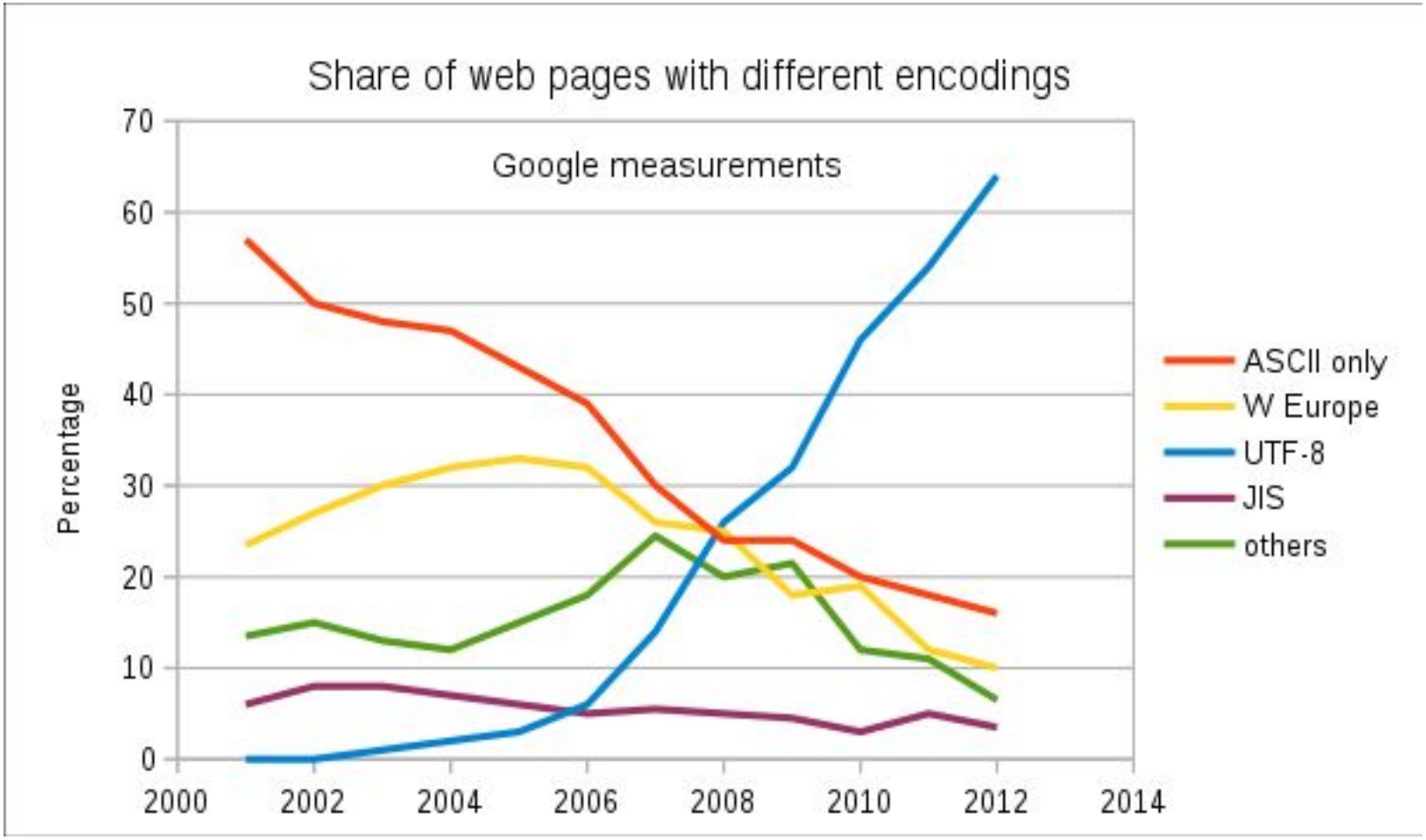
- UTF-8
- UTF-16
- UTF-32

- NOTE: All can express ALL of UNICODE.

Encoding Characteristics

- UTF-8
  - The best! One byte for Ascii! BACKWARD COMPATIBLE!
  - High order bytes never occur in isolation
  - 1-4 characters per symbol.
- UTF-16
  - Good in Asia. 2 or 4 bytes per symbol. (These would typically be 3 bytes per symbol in UTF-8)
  - CJK = Chinese, Japanese, Korean
  - BOM to deal with (big endian vs little endian)
- UTF-32
  - Theoretical mainly, 4 bytes per symbol.

Winner: UTF-8


Share of web pages with different encodings
Google measurements
Percentage
70
60
50
40
30
20
10
0
2000
2002
2004
2006
2008
2010
2012
2014
ASCII only
W Europe
UTF-8
JIS
others

Gotchas

- One Byte != One Character (Symbol)
- Using Unicode doesn't mean using multi-byte characters!
- If working in North America, encode in UTF-8
- If working in Asia, encode in UTF-16
- Encoding is NOT EXPLICIT: often there is no way to denote the encoding. You have to *know* what the encoding is.

In Software

- Carefully Decode. UNICODE. Carefully Encode.
- Beware of “EMOJI BAKE”