Text Encoding

Unicode & UTF

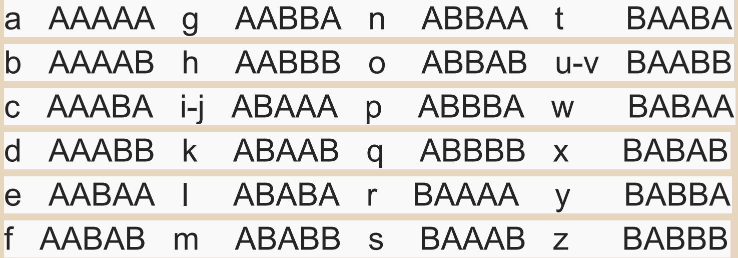
History

One character -> One pattern of encoded bits

Bacon's Cipher

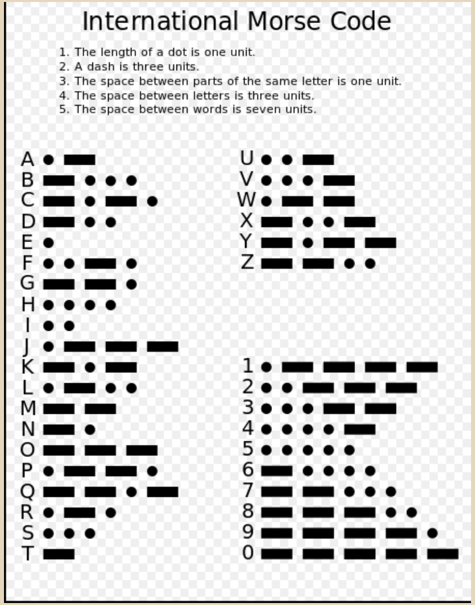
Author: Francis Bacon

Year: 1605



a	AAAAA	g	AABBA	n	ABBAA	t	BAABA
b	AAAAB	h	AABBB	o	ABBAB	u-v	BAABB
c	AAABA	i-j	ABAAA	p	ABBBA	w	BABAA
d	AAABB	k	ABAAB	q	ABBBB	x	BABAB
e	AABAA	l	ABABA	r	BAAAA	y	BABBA
f	AABAB	m	ABABB	s	BAAAB	z	BABBB

Morse Code
Author:
Samuel F.B.Morse
Year: 1836



International Morse Code

1. The length of a dot is one unit.
2. A dash is three units.
3. The space between parts of the same letter is one unit.
4. The space between letters is three units.
5. The space between words is seven units.

A	· −	U	· · −
B	− · · ·	V	· · · −
C	− · − ·	W	· − −
D	− · ·	X	− · · −
E	·	Y	− · − −
F	· · − ·	Z	− − · ·
G	− − ·		
H	· · · ·		
I	· ·		
J	· − − −	1	· − − − −
K	− · −	2	· · − − −
L	· − · ·	3	· · · − −
M	− −	4	· · · · −
N	− ·	5	· · · · ·
O	− − −	6	− · · · ·
P	· − − ·	7	− − · · ·
Q	− − · −	8	− − − · ·
R	· − ·	9	− − − − ·
S	· · ·	0	− − − − −
T	−		

History

- IBM’s Binary Coded Decimal (BCD) - 1959, 6-bit encoding, included: numbers, alphabetic, and special characters.

- ASCII - 1963, 7-bit encoding, included: letters, numerals, symbols, and device control.

- IBM’s Extended Binary Coded Decimal Interchange Code (EBCDIC) - 1963, 8-bit encoding, included: letters, numerals, symbols, and device control

Unicode

Joe Becker (Xerox), Lee Collins (Apple), and Mark Davis (Apple) started researching a universal character set.

- In 1988, Becker first outlined a 16-bit character encoding
- In 1996 Unicode expanded into 21-bit encoding
    - A range of characters U+0000..U+10FFFF
- Unicode can be represented by different Unicode transformation format (UTF)
    - UTF-8
    - UTF-16
    - UTF-32

Important aspects for Unicode

- Code points
- Divided into 17 planes (0 - 16)
    - Each plane has the capacity for 65,536 (=2^16) code points
    - Possibility for 1,114,112 (=65,536 * 17) code points
    - Planes 3-13 are unassigned
    - Basic Multilingual Plane (BMP)
- Surrogates
    - Leading D800 to DBFF
    - Trailing DC00 to DFFF
- Variable-width encoding

UTF-8 vs. UTF-16

UTF-16

- 2 bytes for BMP
- 4 bytes for all other unicode characters
- Big Endian, Little Endian
  - Byte Order Mark (BOM)
  - BE = U+FEFF
  - LE = U+FFFE
- Use surrogates to get full use of plane 1&2

UTF-8 vs UTF-16

- 1 byte for ASCII
- 2 bytes for Arabic, Hebrew, most European Languages.
- 3 bytes for the rest of the BMP
- 4 bytes for all other unicode characters
- Self synchronizing