% Encoding
% Silvestro Di Pietro
% 20/03/2023

Definition

Encoding

In computers, encoding is the process of putting a sequence of characters (letters, numbers, punctuation, and certain symbols) into a specialized format for efficient transmission or storage. Decoding is the opposite process -- the conversion of an encoded format back into the original sequence of characters.

Charsets

Charsets Definition

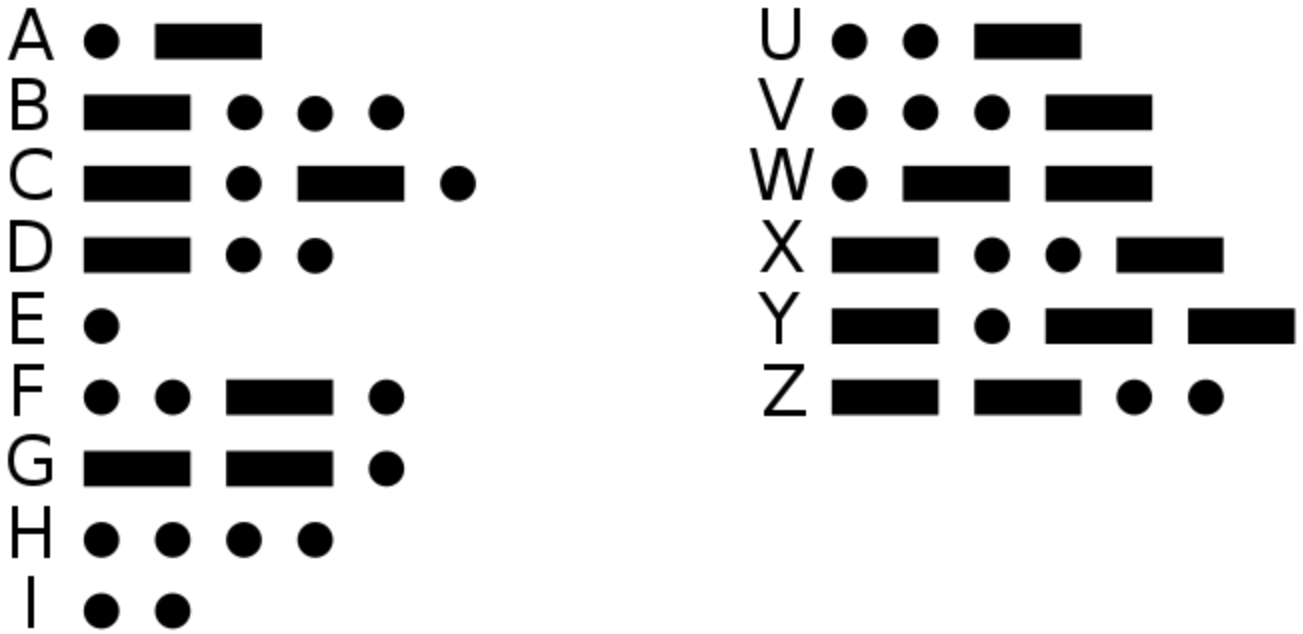
{width=300}

- A character is a minimal unit of text that has semantic value.
- A character set is a collection of characters that might be used by multiple languages

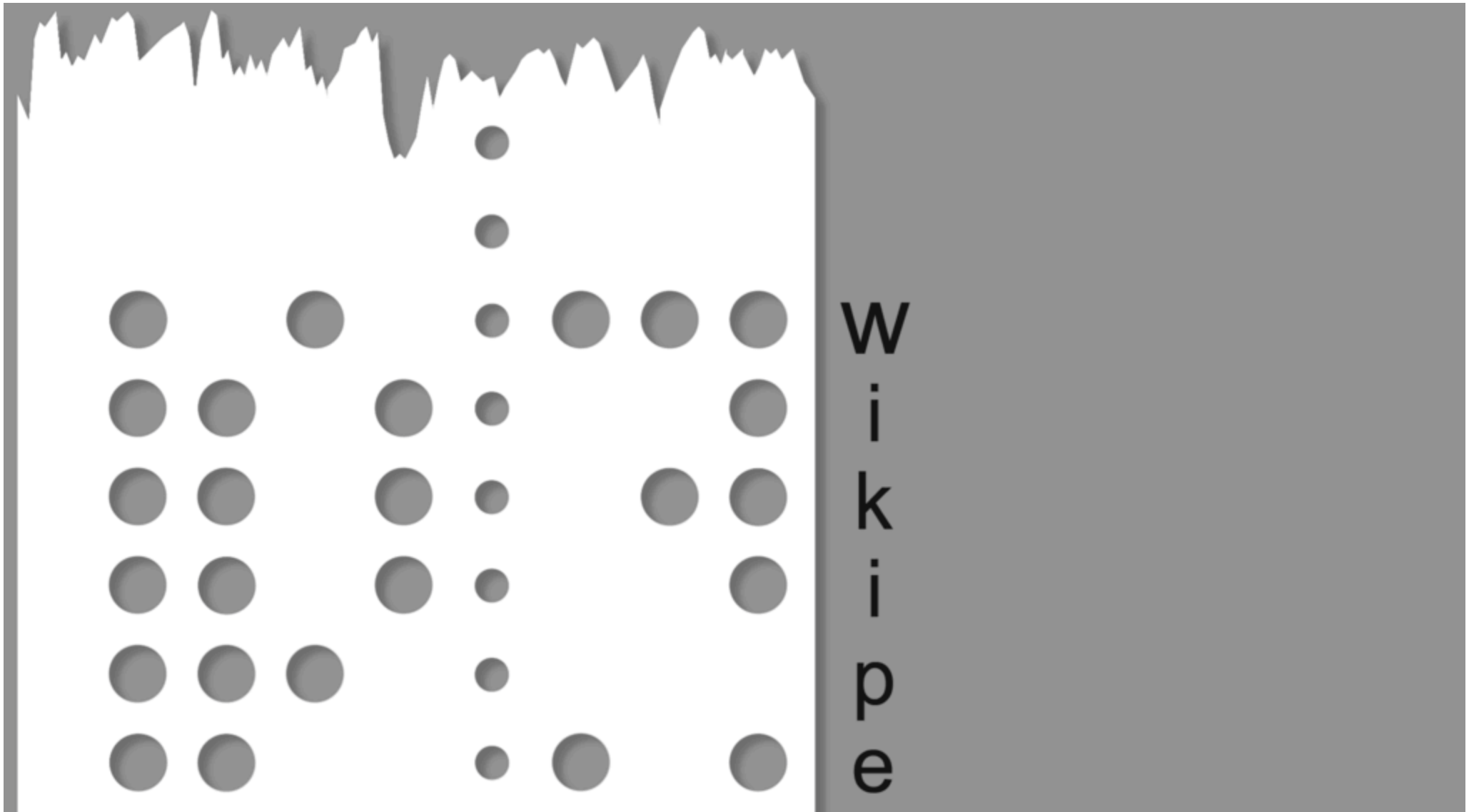
Morse Code

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.



Chars Encoding



Code Unit

A code unit is the "word size" of the character encoding scheme, such as '7-bit', '8-bit', '16-bit'.
In some schemes, some characters are encoded using multiple code units, resulting in a variable-length encoding. A code unit is referred to as a code value in some documents

Code Unit Examples

- A code unit in `US-ASCII` consists of *__7 bits__;
- A code unit in `UTF-8`, EBCDIC and GB 18030 consists of __8 bits__;
- A code unit in `UTF-16` consists of __16 bits__;
- A code unit in `UTF-32` consists of __32 bits__.

Ascii

what is

abbreviated from `American Standard Code for Information Interchange`, is a character encoding standard for electronic communication. ASCII codes represent text in computers, telecommunications equipment, and other devices. Because of technical limitations of computer systems at the time it was invented, ASCII has just `128 code points`, of which only 95 are printable characters.

ASCII table

$b_7 b_6 b_5$ →	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
Bits / Column →	0	1	2	3	4	5	6	7
b_4 b_3 b_2 b_1 / Row ↓								
0 0 0 0	0	NUL	DLE	SP	0	@	P	`	p
0 0 0 1	1	SOH	DC1	!	1	A	Q	a	q
0 0 1 0	2	STX	DC2	"	2	B	R	b	r
0 0 1 1	3	ETX	DC3	#	3	C	S	c	s
0 1 0 0	4	EOT	DC4	$	4	D	T	d	t
0 1 0 1	5	ENQ	NAK	%	5	E	U	e	u
0 1 1 0	6	ACK	SYN	&	6	F	V	f	v
0 1 1 1	7	BEL	ETB	'	7	G	W	g	w
1 0 0 0	8	BS	CAN	(	8	H	X	h	x
1 0 0 1	9	HT	EM	)	9	I	Y	i	y
1 0 1 0	10	LF	SUB	*	:	J	Z	j	z

what is

UTF-8 is a variable-length character encoding standard used for electronic communication. Defined by the Unicode Standard, the name is derived from Unicode (or Universal Coded Character Set) Transformation Format – 8-bit.

UTF8

UTF-8 is capable of encoding all `1,112,064` valid character code points in Unicode using one to `four one-byte (8-bit)` code units.
Code points with lower numerical values, which tend to occur more frequently, are encoded using fewer bytes. It was designed for backward compatibility with `ASCII`:
the first 128 characters of Unicode, which correspond one-to-one with ASCII, are encoded using a single byte with the same binary value as ASCII, so that valid ASCII text is valid UTF-8-encoded Unicode as well.

Conversion: ICONV

The iconv program converts text from one encoding to another encoding. More precisely, it converts from the encoding given for the -f option to the encoding given for the -t option. Either of these encodings defaults to the encoding of the current locale. All the inputfiles are read and converted in turn; if no inputfile is given, the standard input is used. The converted text is printed to standard output.

ICONV example

> man iconv

```
cat exampleISO8859.txt | iconv -f ISO-8859  -t UTF-8 > exampleUTF8
```

UTF8 Example

Character		Binary code point	Binary UTF-8	Hex UTF-8
$	U+0024	010 0100	00100100	24
£	U+00A3	000 1010 0011	11000010 10100011	C2 A3
ह	U+0939	0000 1001 0011 1001	11100000 10100100 10111001	E0 A4 B9
€	U+20AC	0010 0000 1010 1100	11100010 10000010 10101100	E2 82 AC
한	U+D55C	1101 0101 0101 1100	11101101 10010101 10011100	ED 95 9C
𐍈	U+10348	0 0001 0000 0011 0100 1000	11110000 10010000 10001101 10001000	F0 90 8D 88

{width=800}

HTML usage

```
 <!DOCTYPE html>
<html>
        <head>
                <meta charset="utf-8">
        </head>
</html>
```

BASE64

What is

In computer programming, Base64 is a group of binary-to-text encoding schemes that represent **binary data** (more specifically, a sequence of `8-bit` bytes) in sequences of `24 bits` that can be represented by four `6-bit` Base64 digits.

Usage

Binary files cannot be transmitted or stored easily because in a binary file there are bytes, or better, sequence of bits that will interfere within the transmit protocol o the data storage format (eg a binary file sent by email.) or database storage

how encode/decode

. . .

> shell time!

In Unix you can b64 encode a binary file using the shell command `base64`

```
cat images/LogoIfom.png | base64 > images/LogoIfom.b64

cat images/logoIfom.b64 | base64 -d > decoded.png
```

b64 overhead

File name	Size	Base64 size	Base64 gzip size
ifomLogo.png	5754	10233	5533

b64 ENCODING

Index	Binary	Char	Index	Binary	Char	Index	Binary	Char	Index	Binary	Char
0	000000	A	16	010000	Q	32	100000	g	48	110000	w
1	000001	B	17	010001	R	33	100001	h	49	110001	x
2	000010	C	18	010010	S	34	100010	i	50	110010	y
3	000011	D	19	010011	T	35	100011	j	51	110011	z
4	000100	E	20	010100	U	36	100100	k	52	110100	0
5	000101	F	21	010101	V	37	100101	l	53	110101	1
6	000110	G	22	010110	W	38	100110	m	54	110110	2
7	000111	H	23	010111	X	39	100111	n	55	110111	3
8	001000	I	24	011000	Y	40	101000	o	56	111000	4
9	001001	J	25	011001	Z	41	101001	p	57	111001	5
10	001010	K	26	011010	a	42	101010	q	58	111010	6
11	001011	L	27	011011	b	43	101011	r	59	111011	7
12	001100	M	28	011100	c	44	101100	s	60	111100	8
13	001101	N	29	011101	d	45	101101	t	61	111101	9
14	001110	O	30	011110	e	46	101110	u	62	111110	+
15	001111	P	31	011111	f	47	101111	v	63	111111	/

{width=700}

URL encoding

Uniform Resource Identifier

> RFC rfc3986

URL is a subset of th URI and is for Uniform Resource Identifier.
in URL there are some reserved characters as `forward slash`

> ! # $ & ' ( ) * + , / : ; = ? @ [ ]

Percent escaping

You can represent binary chars escaping (prefixing) them using a %

> **A** (see ASCII table) can be represented by `%41`