

# AI6122 Text Data Management & Analysis

Topic: Unicode encoding and UTF-8



# How text is stored

- Computer recognizes and stores **0** and **1**
- How does computer store text and symbols?
  - “Hello World”
  - ☺ ☹
  - “文本处理”
- Encoding scheme
  - Unicode
  - Non-Unicode

$\varepsilon$ 1D700	$v$ 1D710	$E$ 1D720	$Y$ 1D730	$\lambda$ 1D740	$\epsilon$ 1D750	$\Lambda$ 1D760	$\alpha$ 1D770	$\rho$ 1D780
$\zeta$ 1D701	$\varphi$ 1D711	$Z$ 1D721	$\Phi$ 1D731	$\mu$ 1D741	$\vartheta$ 1D751	$M$ 1D761	$\beta$ 1D771	$\varsigma$ 1D781
$\eta$ 1D702	$\chi$ 1D712	$H$ 1D722	$X$ 1D732	$\nu$ 1D742	$\kappa$ 1D752	$N$ 1D762	$\gamma$ 1D772	$\sigma$ 1D782
$\theta$ 1D703	$\psi$ 1D713	$\Theta$ 1D723	$\Psi$ 1D733	$\xi$ 1D743	$\phi$ 1D753	$\Xi$ 1D763	$\delta$ 1D773	$\tau$ 1D783



## Display with different encodings

UTF-8 (8-bit Unicode Transformation Format) 是一种针对Unicode的可变长字符集中的所有有效编码点进行编码，属于Unicode标准的一部分，最初由Ken Thompson提出。它使用8位字节进行编码，这使得原来处理ASCII字符的硬件和软件可以兼容使用。UTF-8就是为了解决向后兼容而设计的编码方式。

A screenshot of a context menu in a file manager. The menu is open over a file named '自2017年1月1日起，所有电子邮件程序都能够使用UTF-8编码形式'. The menu options are: 'Save background as...', 'Set as background', 'Copy background', 'Select all', and 'Paste'. The 'Select all' option is highlighted with a blue background.

UTF-8 尽管如此, 2003年11月UTF-8被RFC 3629 (Unicode 3.1) 定义为“[UTF-8 字节](#)”:

[Add to favorites...](#)  
[View source](#)  
[Inspect element](#)

Encoding	>	Auto-Select
----------	---	-------------

对上述提及的第四种字符而言，UTF-8使用且它的另一种选择，UTF-16编码，对前述视所使用的字符的分布范围而定。不过，如果使用一些传统的压缩系统，比如

**UTF-8**鏺8-bit Unicode Transformation Format鏺銑恹涓?纁?纁?纁?Unicode  
互鑾尤罇鐳冲涑涓? 砢鐳倣 Unicode瀛� 閽囦腑鐳勫壜鏈?爻鎴?鏵堝緞佷俊杩  
**鋒堡鏃** ??鍶?細柄鑼?路媯厠鎖愨語鉅?<sup>[2][3]</sup> 鑾变鈐杈呭富罐鍊肩庡緯杧熒鏗迕罇鐳

[illegible]

菰緯柁	Go to copied address	Ctrl+Shift+L	
鑼?200	Save background as...		相哉菰鏈?涓昏昏 鑼勤紕鑼佺靛寮?紙
剝鼓W	Set as background		鏹?兵卒鑼鏈?變簣鏹?(for all things)鈇
柁熾辜	Copy background		泖鑼鑼豐SCII緯柁熾鏼岃岃涓?哄昂鑼?底I
綉櫻曲	Select all		熾柁妓緯柁熾鏼理紈Shift JIS鋸熾B 23
	Paste		


Consortium, IMC 铸文缓恩 堅鏈委  
ML 鋸函欢鍛子 TML 鋸函欢鑄動規璽よ

UTF-8: Create shortcut  
鑿工師 Add to favorites...

	view source	
1.	Inspect element	𠂇𡙷𢄰𣵀𤕏𥁞Unicode𦛒𧈶𨮓𩆑+

2. Encoding > Auto-Select

Print... Western European (Windows) 湾

3.	Print preview...	Unicode (UTF-8)	
----	------------------	-----------------	---

Refresh	Chinese Traditional (Big5)
	Chinese Simplified (GB18030)

4. Properties

+1FFFF浣跨數鏢溪吐鑼脣紅Unicode

+7FFFFFFF浣跨黠鐳 砵鐳倨級鉅? Right-to-left document

# Unicode

- Unicode is a computing industry standard for the consistent encoding, representation, and handling of text expressed in most of the world's writing systems.
  - The standard is maintained by the Unicode Consortium
  - Unicode 12.1, contains a repertoire of 137,994 characters, covering 150 modern and historic scripts, and multiple symbol sets and emoji.
- Each character is assigned a unique integer code, called “code points”, usually in hexadecimal base
  - A code point is in the form of U+<hex-code>, from U+0000 to U+10FFFF.
  - Characters in English or other languages
  - Currency symbols, Mathematical symbols
  - Emojis e.g., 🐼 U+1F436

bits	character
01000001	A
01000010	B
01000011	C
01000100	D
01000101	E
01000110	F



# UTF-8

- UTF stands for Unicode Transformation Format
- The '8' means it uses 8-bit blocks to represent a character.

1st Byte	2nd Byte	3rd Byte	4th Byte	Number of Free Bits	Maximum Expressible Unicode Value
<b>0</b> xxxxxxx				7	007F hex (127)
<b>110</b> xxxxx	<b>10</b> xxxxxx			(5+6)=11	07FF hex (2047)
<b>1110</b> xxxx	<b>10</b> xxxxxx	<b>10</b> xxxxxx		(4+6+6)=16	FFFF hex (65535)
<b>11110</b> xxx	<b>10</b> xxxxxx	<b>10</b> xxxxxx	<b>10</b> xxxxxx	(3+6+6+6)=21	10FFFF hex (1,114,111)

(Latin Small Letter A With Macron)

ā

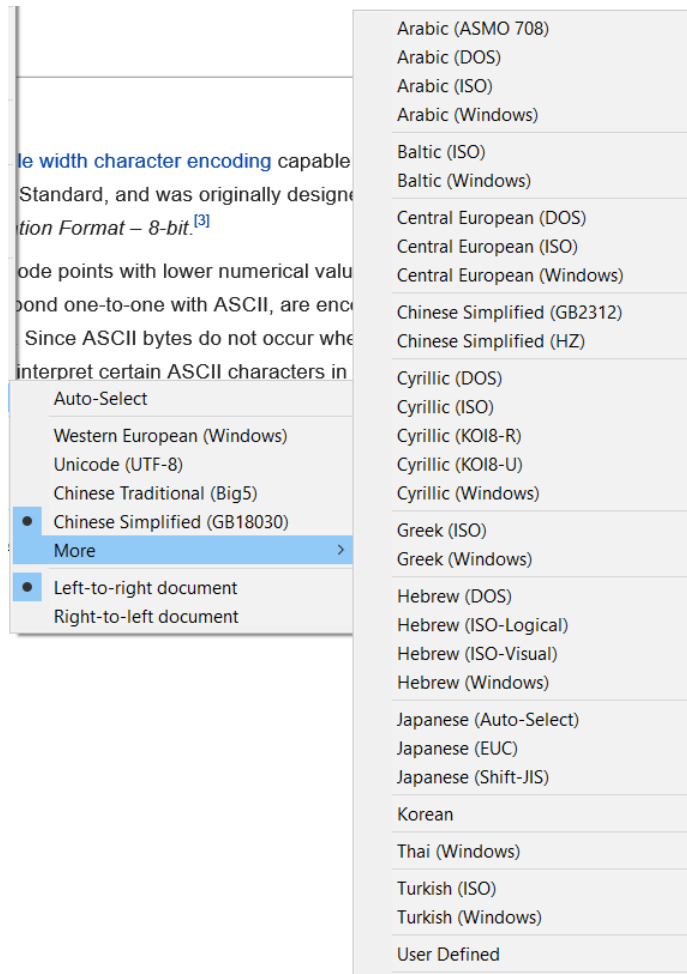
Unicode: decimal 257, binary 100000001

UTF-8 (binary)    **110**00100:**10**000001

<https://www.unicode.org/charts/>



# Unicode Transformation Format (UTF)

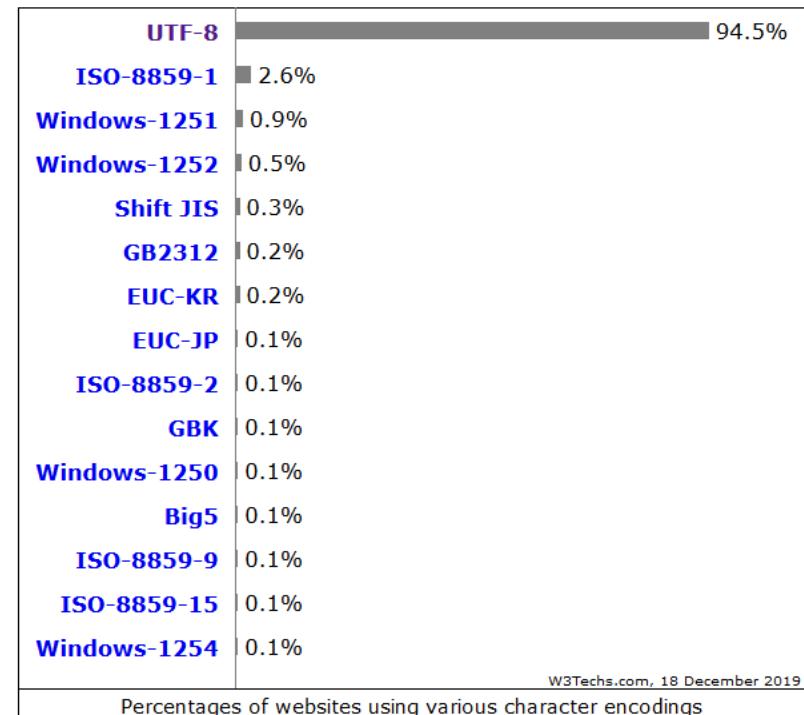


## Usage of character encodings for websites

This diagram shows the percentages of websites using various character encodings [technologies overview](#) for explanations on the methodologies used in the surveys. Reports are updated daily.

How to read the diagram:

UTF-8 is used by 94.5% of all the websites whose character encoding we know.



Source: [https://w3techs.com/technologies/overview/character\\_encoding](https://w3techs.com/technologies/overview/character_encoding)



# Text processing

- Texts are stored in a continuous bit array of 0 and 1s

Hello World

```
01001000 01100101 01101100 01101100 01101111 00100000  
01010111 01101111 01110010 01101100 01100100
```

- Computer does not know any boundary regarding words or sentences
  - There are many different languages
    - With or without explicit word boundaries
    - Reads from left to right or right to left
  - We mainly focus on **English**

