

Unicode

Unicode je standard vyvíjený organizací **Unicode Consortium**, který každému znaku (písmenu, číslici, symbolu) přiřazuje unikátní číselný kód, nezávislý na platformě, programu nebo jazyce. Před vznikem Unicode existovaly stovky různých kódování (např. ASCII, ISO-8859-2 pro češtinu), která byla vzájemně nekompatibilní, což vedlo k chybnému zobrazování znaků (tzv. „rozsypaný čaj“).

Dnešní verze Unicode (v15.1+) definuje více než **150 000 znaků** pokrývajících moderní jazyky, historická písma (egyptské hieroglyfy), matematické symboly a emoji.

Základní principy Unicode

Unicode pracuje se dvěma hlavními vrstvami:

1. Code Point (Kódový bod)

Každý znak má přiřazeno číslo, které se zapisuje v šestnáctkové soustavě s předponou „U+“.

- Například velké písmeno **A** má kódový bod **U+0041**.
- Symbol pro Euro **€** má kódový bod **U+20AC**.
- Emoji pro usměvavou tvář ☺ má kódový bod **U+1F600**.

2. Plane (Rovina)

Prostor Unicode je rozdělen do 17 rovin, z nichž každá pojme 65 536 kódových bodů.

- **BMP (Basic Multilingual Plane)**: Rovina 0, která obsahuje většinu běžně používaných znaků světa.
- Ostatní roviny obsahují méně časté znaky nebo emoji.

Formáty kódování (UTF)

Unicode definuje, jaké číslo má znak mít, ale **UTF** (Unicode Transformation Format) určuje, jak se toto číslo fyzicky zapíše do paměti v [bajtech](#).

UTF-8

Nejpoužívanější kódování na internetu. Je **proměnné délky** (1 až 4 bajty).

- **Zpětná kompatibilita:** Prvních 128 znaků je shodných s ASCII (každý znak zabírá 1 B).
- Česká písmena jako „ž“ zabírají 2 B, asijské znaky 3 B a emoji 4 B.
- **Výhoda:** Extrémně efektivní pro texty v latině.

UTF-16

Znaky ukládá buď jako 2, nebo 4 bajty. Používá se interně v operačních systémech Windows nebo v prostředí Java/.NET.

UTF-32

Každý znak zabírá fixně 4 bajty. Je to neefektivní pro paměť, ale velmi jednoduché pro programátory na výpočet pozice znaku v řetězci.

[Image comparison of UTF-8, UTF-16, and UTF-32 byte storage efficiency]

Unicode a Emoji

Emoji jsou plnohodnotnou součástí standardu Unicode. To je důvod, proč když pošlete „srdíčko“ z iPhone na Android, příjemce jej uvidí (byť v grafice svého systému), protože obě zařízení vědí, že kódový bod **U+2764** znamená srdce.

Zajímavostí jsou **sekvence modifikátorů**, které umožňují měnit barvu pleti nebo kombinovat profese a pohlaví pomocí speciálního znaku **ZWJ** (Zero Width Joiner), který „slepí“ dva kódové body do jednoho výsledného obrázku.

Proč je Unicode důležitý?

| Problém | Řešení pomocí Unicode |
|-----------------------|--|
| Lokalizace | Umožňuje jedné aplikaci běžet v češtině, arabštině i japonštině naráz. |
| Data Integrity | Zabraňuje poškození textu při přenosu mezi různými servery. |
| Vyhledávání | Umožňuje správné řazení a vyhledávání textu bez ohledu na diakritiku. |
| Univerzalita | Podporuje i mrtvá písmena, což je klíčové pro historický výzkum a digitalizaci knih. |

Související pojmy: Byte, ASCII, Šestnáctková soustava, HTML, Browser, JavaScript.

From:

<https://serviceit.cz/> - **IT ENCYKLOPEDIE**

Permanent link:

<https://serviceit.cz/doku.php?id=unicode>

Last update: **2025/12/31 19:33**

