# 2. Digitalizácia textu, ASCII a Unicode, výpočty, porovnanie s praxou

Ľudia pracujú s informáciami. Informácie uložené v počítači sa stávajú údajmi. S údajmi pracujú počítače. Keď potom človek údaje opäť použije pre svoj zámer stávajú sa znova informáciami.
Ak pomocou textového editora píšeme list, na obrazovke ho vidíme v čitateľnom tvare – vidíme a vnímame jeho informačný obsah. Súčasne s touto čitateľnou podobou však program vytvára aj vnútornú podobu nášho listu – binárne zakódovaný tvar dokumentu vhodný pre počítač.
Sú to údaje, ktoré v počítači zodpovedajú nášmu listu. Ak sa rozhodneme list v počítači zachovať, zadáme editoru príkaz Ulož dokument ako ... a vtedy sa údaje, ktoré reprezentujú náš list, zapíšu do súboru.
Je to pomenovaná skupina údajov, ktoré spolu súvisia a sú uložené v počítači na niektorom pamäťovom médiu.

## Digitalizácia textovej informácie:

**1. fáza** – rozdelenia textu na znaky

**2. fáza** - znaky sú pomocou medzinárodne dohodnutej kódovacej tabuľky (v ktorej sú očíslované všetky použité znaky) prekódované na bajty resp. bity.

Každý znak (aj medzera) má svoj kód. Priradenie binárnych kódov znakom sa nazýva **kódová tabuľka**.
Abecedy znakov, ktoré používame pri písaní textu, obyčajne obsahujú viac ako 128 ($2^7$) a menej ako 256 ($2^8$) prvkov. Preto, ako binárny kód znaku použijeme 8 bitov, čiže jeden bajt.
Text sa v počítači zvyčajne ukladá tak, že sa postupne zakódujú jeho znaky.

**Napríklad:**
Meno Janko Zelienka zakódujeme takto:

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Najprv si rozložíme text na znaky: | J | a | n | k | o | | Z | e | l | i | e | n | k | a |
| 2 | Znakom pridelíme číselnú hodnotu podľa tabuľky ASCII: | 74 | 97 | 110 | 107 | 111 | 32 | 90 | 101 | 108 | 105 | 101 | 110 | 107 | 97 |
| 3 | Desiatkové čísla prevedieme do dvojkovej sústavy | 1001010 | 1100001 | 1101110 | 1101011 | 1101111 | 100000 | 1011010 | 1100101 | 1101100 | 1101001 | 1100101 | 1101110 | 1101011 | 1100001 |

Všimnime si, že aj medzera má číselnú hodnotu (32), ktorá sa digitalizuje!

Takže **Janko Zelienka** v "reči počítača" je:
**1001010110000111011101101011110111110000010110101100101111011001101001110010111011101101011110000 1**

Pre veľkosť textového súboru bez formátovacích znakov platí:

veľkosť súboru (v B) = počet znakov x počet bajtov použitého kódu na znak.

Zložitejšie texty obsahujú v počítači aj informácie o svojom formáte, teda o farbe a type písma, o veľkosti strany, veľkosti okrajov a pod.
Tieto informácie sa tiež v digitálnej forme ukladajú do súboru.
Najrozšírenejší kód na kódovanie znakov je **ASCII** (American Standard Code for Information Interchange) kód. Je 7 bitový, teda umožňuje zakódovať iba 128 znakov.
Pridaním jedného bitu umožňuje zakódovať aj znaky národných abecied, takto upravený sa nazýva **rozšírený ASCII kód**. Znaky jednotlivých národných abecied tvoria kódové stránky. **Každá národná abeceda má svoju kódovú stránku.**

8 bitový, t.j. 1 znak je zakódovaný v 1 bajte; 8 bitov umožňuje zakódovať 28 = 256 rôznych znakov;

prvých 128 znakov je pevne daných (z písmen len anglická abeceda; stačil by 7 bitový kód),

zvyšných 128 znakov sa mení podľa nastaveného jazyka (národné abecedy).

Tabuľka základných znakov ASCII (128 znakov)

| kód / znak | | kód / znak | | kód / znak | | kód / znak | | kód / znak | | kód / znak | | kód / znak | | kód / znak | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | 16 | | 32 | | 48 | 0 | 64 | @ | 80 | P | 96 | ` | 112 | p |
| 1 | | 17 | | 33 | ! | 49 | 1 | 65 | A | 81 | Q | 97 | a | 113 | q |
| 2 | | 18 | | 34 | " | 50 | 2 | 66 | B | 82 | R | 98 | b | 114 | r |
| 3 | | 19 | | 35 | # | 51 | 3 | 67 | C | 83 | S | 99 | c | 115 | s |
| 4 | | 20 | | 36 | $ | 52 | 4 | 68 | D | 84 | T | 100 | d | 116 | t |
| 5 | | 21 | | 37 | % | 53 | 5 | 69 | E | 85 | U | 101 | e | 117 | u |
| 6 | | 22 | | 38 | & | 54 | 6 | 70 | F | 86 | V | 102 | f | 118 | v |
| 7 | | 23 | | 39 | ' | 55 | 7 | 71 | G | 87 | W | 103 | g | 119 | w |
| 8 | | 24 | | 40 | ( | 56 | 8 | 72 | H | 88 | X | 104 | h | 120 | x |
| 9 | | 25 | | 41 | ) | 57 | 9 | 73 | I | 89 | Y | 105 | i | 121 | y |
| 10 | | 26 | | 42 | * | 58 | : | 74 | J | 90 | Z | 106 | j | 122 | z |
| 11 | | 27 | | 43 | + | 59 | ; | 75 | K | 91 | [ | 107 | k | 123 | { |
| 12 | | 28 | | 44 | , | 60 | < | 76 | L | 92 | \ | 108 | l | 124 | \| |
| 13 | | 29 | | 45 | - | 61 | = | 77 | M | 93 | ] | 109 | m | 125 | } |
| 14 | | 30 | | 46 | . | 62 | > | 78 | N | 94 | ^ | 110 | n | 126 | ~ |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 15 | | 31 | | 47 | / | 63 | ? | 79 | O | 95 | _ | 111 | o | 127 | □ |

**Univerzálny kód**, ktorý umožňuje zakódovať rôzne znaky zo všetkých bežných jazykov, sa nazýva **UNICODE**. Je 16 bitový a obsahuje $2^{16}$ = 65 536 znakov.

16 bitový, t.j. 1 znak je zakódovaný v 2 bajtoch; 16 bitov umožňuje zakódovať 216 = 65 536 rôznych znakov; pozri napr. Word 2002 *Vložiť – Symbol... – Symboly* - Písmo: (normálny text)

Príklad: V programe Word otvorte jednoduchý textový dokument (bez obrázkov, tabuliek, hlavičky a päty a pod.) a cez *Súbor – Vlastnosti – Štatistika* - Štatistika: Znakov (vrátane medzier) zistite počet znakov v texte. Text prekopírujte do aplikácie Poznámkový blok a uložte ako textový súbor (*.txt). Presvedčte sa, že veľkosť súboru je len málo väčšia (pri jednoriadkovom súbore priamo zodpovedá) súčinu: počet znakov krát bajt.

Aký kód je použitý v textovom súbore? Ako by sa zmenila veľkosť súboru pri použití druhého kódu?

Pre veľkosť textového súboru bez formátovacích znakov platí:

**veľkosť súboru [v B] = počet znakov x počet bajtov použitého kódu na znak [B/znak]**

Ak by boli v texte vytvorené odseky, každé stlačenie klávesu Enter je v textovom súbore zakódované dvoma znakmi – CR (Carriage Return/na začiatok riadka), riadiaci kód 1310 (D16) a LF (Line Feed/nový riadok), riadiaci kód 1010 (A16).

## Otázky a úlohy

1. Čo je to digitalizácia text?
2. Aký je rozdiel medzi ASCII a Unicode tabuľkou?
3. Vysvetli ako vypočítame veľkosť textového súboru bez formátovania?
4. Vysvetlite ako vyzerá kódová tabuľka ASCII (koľko znakov obsahuje, koľkými bitmi zapisuje jednotlivé znaky, aké znaky našej abecedy tam nie sú).
5. Vysvetlite ako vyzerá kódová tabuľka Unicode (koľkými bitmi sa zapisuje každý znak, koľko znakov umožní zapísať táto tabuľka, koľko pamäte zaberú znaky v porovnaní s kódovaním ASCII).