

# Text

Textová data jsou data, která obsahují informaci vyjádřenou pouze slovy a větami nějakého jazyka – přirozeného nebo formálního.

Text v přirozeném jazyce je určen ke čtení lidmi.

Text ve formálním jazyce (programovacím, konfiguračním, datovém) je určen ke čtení programem, ale vždy je také bez převodu čitelný člověkem.

Mezi textová data *nepatří* spustitelné programy ve strojovém kódu, obrázky, zvuk, video, ani binární (netextová) data určená aplikačním nebo systémovým programům.

# Prostý a formátovaný text

Text může být

- *prostý*, který je neformátovaný a obsahuje jen základní znaky v jediné velikosti a řezu,
- *formátovaný*, který vznikne z prostého textu jeho obohacením o *označkování*
  - *logickými značkami* upřesňujícími a rozšiřujícími význam některých částí textu
  - a *fyzickými značkami* popisujícími formát a podobu částí textu.

# Prostý text

Prostý text obsahuje jen

- písmena a á b c č A À Ç ...
- číslice 0 1 2 ...
- interpunkci . , : ; ? ! [ ] ( ) / ...
- speciální zobrazitelné znaky # % @ ...
- a *bílé znaky* – mezeru a oddělovač řádků.

Písmo je omezeno na jedinou velikost, řez a duktus.

Často má jednotnou šířku všech znaků, neboli je *neproporcionální*.

Prostý text neobsahuje ani *řídící znaky*, s jedinou výjimkou – znaku *oddělovače řádků*.

# Proporcionální a neporcionální písmo

V prostém textu jsou řádky stejně daleko od sebe, nerozlišují se druhy písma ani jeho velikost.

Při zobrazování prostého textu se často používá *neporcionální* písmo, též zvané *strojopisné*. Znaků tohoto písma mají všechny stejnou šířku.

## Příklad:

|                |                                  |       |        |
|----------------|----------------------------------|-------|--------|
| Proporcionální | – šířka různých znaků je různá:  | Jiljí | MBWEHA |
| Neporcionální  | – šířka různých znaků je stejná: | Jiljí | MBWEHA |

# Neproporcionální písmo

Neproporcionální písmo se používá v *textovém uživatelském rozhraní* operačního systému. Dříve se používalo na *mechanických psacích strojích*. Když je použijeme při zobrazení textu, zdůrazňujeme tím, že text je prostý a neformátovaný.

Text v neproporcionálním písmu se používá i uvnitř formátovaného textu pro jména složek, souborů, systémových služeb, systémových a aplikačních programů, textových výstupů, chybových hlášení.

Tím typograficky vyjadřujeme, že se jedná o přesný tvar jmen a příkazů z textové komunikace s operačním systémem nebo s aplikačním programem.

# Textové editory

Soubory s prostým textem mívají jména s příponou `.txt` nebo `.text`.

Pořizujeme je pomocí *textových editorů*, například

- atom,
- emacs,
- vim,
- nano,
- notepad.

# Znaková sada

Množina znaků, které se používají v určitém jazyce nebo ve skupině jazyků. Typicky obsahuje *písmena*, *číslice*, *interpunkční znaménka* a další (zobrazitelné) znaky a symboly. Může obsahovat i speciální *řídící* znaky.

## Příklady:

*ASCII* je znaková sada pro angličtinu; 95 zobrazitelných + 33 řídících znaků.

*Latin-1 (ISO-8859-1)* sada pro západoevropské jazyky; 256 znaků

*Latin-2 (ISO-8859-2)* sada pro středo- a východoevropské jazyky; 256 znaků

*CP1250* jiná sada pro středo- a východoevropské jazyky

*Unicode* sada pro všechny známé jazyky; skoro 150 000 znaků.

Pozor, znaková sada *není* kódování. Jedna znaková sada může mít různá kódování.

# Kódování znaků

Kódování znakové sady určuje, jak se každý znak vyjádří v počítači pomocí bitů. Kromě toho může kódování předepisovat pořadí těchto bitů v bytu nebo slově (tzv. *endianitu*). Kódování textu bývá doplněno informací o způsobu oddělování řádků.

U malých znakových sad je kódování znaků přímočaré: každý znak má své číslo (ASCII od 0 do 127, Latin-x od 0 do 255) a to se zapíše jako celé číslo beze znaménka ve dvojkové soustavě do jednoho bytu.

Sada Unicode má více kódování. Například *UCS-32* má každý znak přesně ve 32 bitech. V kódování *UTF-8* mají znaky různě dlouhé kódy: 8, 16, 24 nebo 32 bity. Pro sadu Unicode existují i další kódování: *UCS-16*, *UTF-16*...

Pro převody textů do různých kódování slouží k tomu určené aplikační programy, například *iconv* nebo *DenCode*. Lze použít i nástroje online.

## Oddělování řádků v textu

Sousední řádky textu bývají od sebe odděleny jedním nebo dvojicí řídicích znaků. Různé skupiny operačních systémů mají různé konvence.

|       |           |                                                        |
|-------|-----------|--------------------------------------------------------|
| LF    | 0x0A      | unixové operační systémy – Linux, macOS, FreeBSD . . . |
| CR LF | 0x0D 0x0A | MS-Windows, DOS, PalmOS                                |
| CR    | 0x0D      | starší verze MacOS na počítačích Apple                 |

Oddělovač řádků slouží i jako *ukončovač* řádku. V souboru s prostým textem by měl být i za posledním řádkem, tj. jako poslední byte souboru.

## Oddělování řádků v textu

Při přenosu textových dat mezi počítači se o správný převod starají aplikační programy podle síťových protokolů.

Při ručním přenosu souborů však musíme dát na oddělování řádků pozor, jinak se text zobrazí nesprávně.

**Příklad:** Posíláme text poštou mezi počítači s operačními systémy s různými konvencemi pro oddělování řádků.

- Když text napíšeme do těla zprávy, klientské poštovní programy a poštovní protokol se samy postarají o správný převod kódování i oddělování řádků.
- Když soubor s textem zkomprimujeme a pošleme jako přílohu, poštovní programy do něj nezasahují a příjemce musí kódování i oddělování řádků ošetřit ručně.

## Cvičení:

- V textovém editoru napište nějakou oblíbenou básničku nebo jiný krátký text.
- Text uložte do souboru na sdíleném disku (např. K:basnicka.txt).
- Zjistěte, jak vypadají oddělovače řádků.

```
hd, hexdump, get-content -encoding byte
```

- Zobrazte text v terminálovém okně

```
cd K:  
cat basnicka.txt
```

- Pokud se text nezobrazil správně, převedte ho do kódování, které používá textové okno  
iconv nebo pspad

## Cvičení:

- V souboru basnicka.txt odstraňte ukončení posledního řádku.
- Vypište soubor v textovém okně. Co pozorujete?

# Formátovaný text

Formátovaný text obsahuje navíc fyzické nebo logické značky.

Má dvě podoby:

- *cílový tvar* zformátovaný podle značek, připravený k tisku nebo finální prezentaci.  
Textové procesory typu WYSIWYG zobrazují zformátovaný dokument v cílovém tvaru.
- *zdrojový tvar* obsahující text spolu se značkami;  
Zdrojový tvar se zviditelněnými značkami je vidět v textových editorech.  
Například v souborech .odt, .html apod. Formátovací značky však mohou být i netextové; editor je pak zobrazuje jako neinterpretované řídicí znaky.

## Příklad:

```
<h3>Nadpis</h3>  
Text obsahující <i>kursívu</i>  
a také <b>tučné písmo</b>.
```

Zdrojový tvar — prostý text se značkami

## Nadpis

Text obsahující *kursívu* a také **tučné písmo**.

Cílový tvar — formátovaný text

# Označovaný text

*Zdrojový tvar* formátovaného textu obsahuje *označkování*.

V cílovém tvaru už značky nevidíme přímo, textový procesor nebo sázecí systém je interpretoval formátováním textu.

Značky se

- buď jen přidají do prostého textu a ve zdrojovém je vidíme jako značky (*tags*) se zvláštní syntaxí; uživatel vidí zdrojový i cílový tvar textu.

HTML: text `<i>kursívou</i>` nebo `<b>tučně</b>`

T<sub>E</sub>X: text `\textit{kursívou}` nebo `\textbf{tučně}`

MediaWiki: text `'kursívou'` nebo `'''tučně'''`

Markdown: text `*kursívou*` nebo `**tučně**`

- anebo je dokumentový procesor vkládá do binárního souboru, který není čitelný přímo, ale jen po překladu do cílového dokumentu; uživatel vidí jen cílový (zformátovaný) tvar textu.

Takto funguje například MS Office.

# Fyzické značkování

Fyzicky označovaný (marked-up) text obsahuje značky, které určují, jak má daná část textu vypadat po převedení do cílové podoby: zobrazení na monitoru, vytištění na papír, případně po exportu do PDF či jiného grafického formátu.

Fyzické značky určují

- písmo a jeho velikost, řez, duktus, barvu popředí a pozadí,
- rozměry stránky a tiskového zrcadla,
- řádkový rozpal, vzdálenost mezi odstavci,
- počet sloupců sazby,
- rozměry a tvar tabulek, barvy a tloušťky čar,
- odsazování, atd.

# Logické značkování

Logické značkování určuje role částí textu a vztahy mezi částmi textu.

Například

- hranice celků: kapitoly, podkapitoly, odstavce,
- strukturované útvary v textu: výčty, seznamy, tabulky,
- nadpisy různých úrovní,
- zdůrazněné části,
- poznámky,
- vložené netextové objekty: obrázky, diagramy, videosekvence.

# Styl dokumentu

V delších textech se nedoporučuje používat přímo fyzické značky.  
Vhodnější je označkovat text logicky a styl dokumentu definovat zvlášť.  
Tím dosáhneme

- jednotné podoby cílového dokumentu
- jednodušší změny podoby dokumentu — změny se dělají na jednom místě
- jednoduššího generování dokumentů
- možnosti mít více stylů pro jeden dokument
- možnosti mít styl společný více dokumentům

# Styl dokumentu

Dokument, který je označován čistě logicky (bez fyzických značek), sám neřeší vnější podobu cílového textu. Ta se řeší odděleně v takzvaném *popisu stylu*.

Styl přiřazuje logickým značkám příslušné formátování (kombinaci fyzických značek). Formátování konkrétního logického celku označeného logickou značkou závisí

- na logické značce  
číslovaný seznam bude vypadat jinak než tabulka
- na jejím umístění v textu  
zvýrazněný text uvnitř zvýrazněného textu bude vysázen jinak než okolní text
- na typu dokumentu  
odstavec v knize bude mít jiné odsazení než v krátkém článku
- na parametrech dokumentu, například rozměrech stránky  
při větší šířce může být sazba vícesloupcová

# Značkovací jazyky

Existuje mnoho systémů značkování. Často kombinují fyzické značkování s logickým a je ponecháno na uživateli, zda chce logickou strukturu dokumentu oddělit od stylu.

**Příklad:** Značkovací jazyk HTML užívá směs fyzického i logického značkování. Například fyzické `<i>`, `<b>`, `<br>`; logické `<body>`, `<h3>`, `<em>`, `<strong>`, `<p>`.

Nová verze, jazyk HTML5, dovoluje už jen logické značky. Vzhled stránek se definuje odděleně, pomocí stylového souboru CSS (*cascading style sheet*).

## Příklady značkovacích jazyků

| značkování          | použití                                                                                                                      | nástroje<br>pro pořízení                       | nástroje<br>pro zobrazení      |
|---------------------|------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------|--------------------------------|
| HTML                | Pro webové stránky. Značky jsou textové, takže lze použít textový editor, ale je možný i export z jiného značkování do HTML. | textové editory<br>(i s podporou HTML)         | prohlížeče html<br>HTML5 Maker |
| EPUB                | Pro e-knihy.                                                                                                                 | Adobe, Calibre, Okular                         |                                |
| troff               | Určený pro formátování informačních stránek v unixových systémech (man pages). Značky jsou textové.                          | textové editory                                | man, troff                     |
| ODT, DOC            | Obecné formátované dokumenty určené k tisku, promítání i zobrazení na monitoru.                                              | nástroje skupiny Office: LibreOffice, Word ... |                                |
| T <sub>E</sub> X    | Kvalitní sazba určená k publikaci knih a vědeckých článků. Značky jsou textové.                                              | textové editory                                | prohlížeče PDF<br>Overleaf     |
| Markdown, MediaWiki | Dokumenty přehledné a snadno čitelné i ve zdrojovém tvaru                                                                    | textové editory                                |                                |