

# Kódování

- co to je kódování (obecně)
- komprese dat
- samodetekující kód (příklady)
- přenosové kódování (MIME)
- kódové stránky (jednobytové a vícebytové)
- kodek

## Kódování

- záznam informace pomocí pevně stanovené znakové tabulky nebo znakové sady (např. melodie je zapsaná v notách)
- kódování často znamená také převod již kódované informace do jiného kódu, jako je například transliterace textu do jiné abecedy nebo převod elektronické informace do jiného kódu či normy
- zvláštní význam má převod „otevřené“ informace do kódu, který je znám jen určitým osobám, čili [šifrování](#) (kryptografie)
- převod kódované informace do obecně přístupného tvaru je **dekódování** nebo dešifrování.

## Komprese dat

- speciální postup, při ukládání nebo transportu dat
- úkolem komprese dat je zmenšit datový tok
- vhodné např. pro archivaci nebo pro přenos dat přes síť s omezenou rychlostí (snížení doby nutné pro přenos)
- komprese může být také nutná při omezené datové propustnosti (mobilní telefony komprimují hovor pro přenos přes GSM síť)

*Lze rozdělit do 2 základních kategorií:*

### • komprese ztrátová

- některé informace jsou nenávratně ztraceny a nelze je zpět rekonstruovat
- používá se tam, kde je možné ztrátu některých informací tolerovat a kde nevýhoda určitého zkreslení je bohatě vyvážena velmi výrazným zmenšením souboru
- používá se pro kompresi zvuku a obrazu, kde si člověk do určité míry chybějících údajů nevšimne nebo si je dokáže domyslet (je využíváno nedokonalostí lidského zraku a sluchu)
- např. JPEG, MPEG, MP3, ...

### • komprese bezztrátová

- obvykle není tak účinná jako ztrátová komprese
- komprimovaný soubor lze opačným způsobem rekonstruovat do původní podoby (ale ztráta jediného znaku může znamenat nenávratné poškození souboru)
- např. FLAC, GIF, PNG, soubory ZIP, RAR, apod.

- **Kompresní poměr** je podíl velikosti původních dat ku velikosti dat komprimovaných
  - Při kompresi 10MB souboru do 2MB souboru je kompresní poměr 5 : 1 (pětkrát zmenšeno), úspora je tedy 80%
  - Kompresní poměr je ovlivněn volbou kompresního algoritmu i typem komprimovaných dat
  - Například nekomprimované skladby na audio CD mají datový tok přibližně 1,35Mb/s, zatímco komprimované zvukové soubory (MP3, AAC) mají datový tok 128Kb/s

## Samodetekující kódy

- slouží k jednoduchému ověření správnosti zadaných dat
- princip je založen na tom, že zadaná data jsou (buď rovnou, nebo po určité matematické operaci) dělitelná daným číslem, nejčastěji bývá voleno 11 (tzv. *jedenáctkový samodetekující kód*)
- **daný dělitel by měl splňovat několik kritérií:**
  - dvoucifernost (aby se dala vždy odhalit chyba zadaná v jedné cifře)
  - prvočíselnost
- čím vyšší dělitel se zvolí, tím větší bude pravděpodobnost odhalení chyby (ale o to delší kód)
- příklady
  - EAN-13 (čárový kód)
  - rodná čísla vytvořená od roku 1986 (jsou dělitelná jedenácti - poslední čtyři číslice jsou voleny tak, aby každé rodné číslo tuto vlastnost splňovalo)
  - ISBN (jedinečná identifikace knižního titulu), ISSN (jedinečná identifikace periodických publikací - noviny, časopisy, včetně těch vycházejících online)
  - čísla bankovních účtů a kreditních karet
- **nedostatky**
  - je možné ověřit správnost zadání, při chybě ale není možné zjistit původní informaci
  - tento nedostatek bývá odstraněn pomocí redundance informace u samoopravných kódů

Redundance znamená informační nebo funkční nadbytek, například větší množství informace, prvků nebo zařízení než je nezbytné. Zejména zápisy čísel, kódů a programů mají velmi nízkou redundanci, která se často plánovitě zvyšuje například paritou, kontrolní číslicí nebo kontrolním součtem, které umožňují aspoň odhalení části chyb. Ještě daleko složitější a nákladnější redundance se užívají v podobě tzv. „samoopravného kódování“ (samodetekující kódy), které dovoluje automatickou opravu jedné nebo i více chyb.

## Přenosné kódování (MIME)

- internetový standard, jedná se o rozšíření internetové pošty
- umožňuje:
  - zasílat zprávy s diakritikou (podpora textu psaného ve znakových sadách jiných než US-ASCII)
  - podpora příloh (obrázky, zvuky, atp.)
  - vícedílné zprávy
  - využít funkci digitálního podpisu
  - informace v hlavičce v jiné znakové sadě než ASCII
- v současné době používán i dalšími protokoly (např. HTTP)

## Kódové stránky (jednobytové a vícebytové)

Kódová stránka (znaková sada nebo také kódování) je kód, který páruje sekvence znaků z dané množiny (abecedy) s jejich jinou reprezentací, jako je sekvence přirozených čísel, bajtů nebo elektrických pulzů, za účelem ukládání textu v počítači nebo přenosu textu telekomunikačními sítěmi. Kódová stránka např. říká, že bajt s číselnou hodnotou **65** má být považován za písmeno velké **A**.

Může se jednat např. o Morseovu abecedu, v prostředí počítačů o znakovou sadu **ASCII**. Kódování je také prostředkem pro kompresi (tj. zmenšení) anebo šifrování (tj. utajování) dat.

- **ASCII** - standard z roku 1963, kóduje 128 znaků americké abecedy (společně s číslicemi a dalšími symboly) jako 7bitová čísla
  - jedná se o základní kódování, z kterého vychází v euro-americkém prostoru ostatní standardy
  - drtivá většina osmibitových (jednobytových) znakových sad pouze rozšiřuje ASCII tím, že přidávají významy kódům 128-255, které se v ASCII nepoužívají
- **8-bitová kódování** - jeden byte odpovídá jednomu konkrétnímu znaku (byte s hodnotou 65 = písmeno velké A)
  - 8-bitová kódování pro češtinu
    - **Windows-1250** (někdy také označován jako CP-1250; na platformě Windows)
    - **ISO 8859-2** (UNIXové systémy)
    - **Kód Kamenických** (kódování částečně kompatibilní s CP437 (zachovává semigrafické znaky))
  - 8-bitová kódování nejsou schopna pojmout znaky některých abeced, které obsahují víc jak 255 znaků (např. čínština, apod.)
- **Vícebytová kódování** - jednomu znaku odpovídá sekvence více bajtů
  - takováto kódování (především standard Unicode) se používají, kvůli možnosti uložit a zobrazit písmena všech myslitelných abeced v rámci jedné znakové sady
  - díky vícebytovým kódováním lze např. v rámci jednoho textového souboru přirozeně kombinovat znaky anglické, české a japonské abecedy
  - to je hlavní rozdíl od 8-bitových kódování, u kterých je možné v rámci dokumentu použít jenom omezený počet znaků (max. 255)
  - **Unicode** - tabulka znaků všech existujících abeced, která v současnosti obsahuje několik stovek tisíc znak
    - každý znak má jednoznačný číselný kód a svůj název
    - definuje u každého znaku některé základní vlastnosti jako např. zda se jedná o písmeno, symbol atd., zda je písmeno velké či malé atp.
    - prvních 128 znaků (tj. sedmibitové kódy) obsahuje znakovou sadu ASCII
    - Unicode se kóduje různými způsoby - mezi základní metody patří:
      - UTF-32 (každý znak reprezentován přímo 32bitovým číslem)
      - UTF-16 (každý znak reprezentován jedním 16bitovým číslem, znaky, které není možné uložit jako 16bitové číslo, jsou reprezentovány párem 16bitových čísel)
      - UTF-8
  - **UTF-8**
    - každý znak má jinou délku (1-4 bajty)

- prvních 128 znaků (U+0000 – U+007F (0-127)) je zapsáno pomocí 1 bajtu (odpovídají ASCII kódování)
- jedná se o výchozí kódování pro **XML** dokumenty (v případě, že neurčíme jinou znakovou sadu)
- **Výhody**
  - používá se hlavně pro přenos dat, neboť je **prostorově úsporné** (znaky běžných písem jsou kódovány krátkými posloupnostmi; nevýhodu mají uživatelé písem Dálného východu)
  - odolné proti chybám
  - zpětně kompatibilní s ASCII
- **Nevýhody**
  - odlišné délky jednotlivých znaků

## Kodek

- zkratka dvou slov **KO**der + **DEK**oder (respektive **kompre**se, **dek**omprese)
- zařízení nebo počítačový program, který dokáže transformovat datový proud nebo signál
- software, který se používá ke kompresi a dekompresi digitálního mediálního souboru, například skladby nebo filmu. Program Windows Media Player a další programy používají kodeky k přehrávání a vytváření digitálních mediálních souborů
- počítačový program nebo hardwarové zařízení, které kóduje a dekóduje video do/z určitého formátu, zpravidla za účelem zmenšení objemu dat.; takový video proud se pak běžně ukládá do tzv. multimediálního kontejneru, který umožňuje kombinovat různé multimediální datové proudy (audio, video, titulky) do jednoho souboru
- kodeky se dají rozdělit několika způsoby; základním je dělení na *bezeztrátové* a *ztrátové*
  - **bezeztrátové:**
    - *Huffyuv* – využívá Huffmanova kódování, které je založeno na různé četnosti znaků
      - výhodou je rychlá komprese a dekomprese
  - **ztrátové:**
    - *DivX*
      - používá standardní kompresi MPEG-4 ASP, nikoli vlastní formát videa (je kompatibilní s ostatními MPEG-4 ASP kodeky)