ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
FAKULTA STAVEBNÍ

OBOR GEODÉZIE A KARTOGRAFIE
KATEDRA MAPOVÁNÍ A KARTOGRAFIE

Konverze textových formátů

semestrální práce

Martin Setnička

Michal Šatava

V Praze dne 29. 3. 2009

Kartografická polygrafie a reprografie

Úvod

Následující text by měl především vyložit základní pojmy z oblasti vyjadřování textu v elektronické podobě, pojednat o problémech při čtení a konverzi textových formátů a nastínit cestu, kterou se ubírat při jejich řešení.

1 Základní pojmy

Terminologie je dosti nejednoznačná, proto uvádím krom základních pojmenování i ostatní jména v závorce.

Dokument

Soubor obsahující *vlastní text* a *formátovací značky*. Podle zápisu obsahu se jedná o *textový* nebo *binární formát*. Formát je popsán příponou a nebo v záhlaví souboru.

Textový formát

Každému *znaku* (symbolu) je přiřazeno celé nezáporné číslo, podle dané *znakové sady* (kódová tabulka, kód). Toto číslo je podle *kódovacího schématu* (charset) zapsáno n celými byty (podle velikosti znakové sady). Nezabývá se vizuální prezentací znaku (*glyf*)

Např: HTML, XML, PostScript, TeX, CPP, RTF, CSV, SVG, ...

Binární formát

Alespoň část informací je vyjádřena jinak než v textovém formátu.

Např: DOC, ODT, Text602, ...

2 Textový formát

2.1 Jednobytové znakové sady

2.1.1 US-ASCII

ASCII je zkratka pro American Standard Code for Information Interchange (americký standardní kód pro výměnu informací).

Jde o první znakovou sadu. Sedm bitů definuje znaky anglické abecedy a jiné znaky používané v informatice. Osmý bit je nulový. Obsahuje tedy 128 znaků (viz Obr. 1).

2.1.2 8-bitová kódování (extended text)

Rozšíření využívající osmého bitu US-ASCII kódu, která obsahují dalších 128 kódů. Jsou bohužel i případy, kdy prvních 128 znaků není totožných s US-ASCII.

USASCII code chart

Bits b_7 b_6 b_5 →	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
b_4 b_3 b_2 b_1 ↓ Row / Column →	0	1	2	3	4	5	6	7
0 0 0 0	0	NUL	DLE	SP	0	@	P	`	p
0 0 0 1	1	SOH	DC1	!	1	A	Q	a	q
0 0 1 0	2	STX	DC2	"	2	B	R	b	r
0 0 1 1	3	ETX	DC3	#	3	C	S	c	s
0 1 0 0	4	EOT	DC4	$	4	D	T	d	t
0 1 0 1	5	ENQ	NAK	%	5	E	U	e	u
0 1 1 0	6	ACK	SYN	&	6	F	V	f	v
0 1 1 1	7	BEL	ETB	'	7	G	W	g	w
1 0 0 0	8	BS	CAN	(	8	H	X	h	x
1 0 0 1	9	HT	EM	)	9	I	Y	i	y
1 0 1 0	10	LF	SUB	*	:	J	Z	j	z
1 0 1 1	11	VT	ESC	+	;	K	[	k	{
1 1 0 0	12	FF	FS	,	<	L	\	l	|
1 1 0 1	13	CR	GS	-	=	M	]	m	}
1 1 1 0	14	SO	RS	.	>	N	^	n	~
1 1 1 1	15	SI	US	/	?	O	_	o	DEL

Obr. 1: Tabulka ASCII kódů [3]

Rozšířený kód je přesto malý na to, aby pojmul třeba jen evropské národní abecedy. Vzniklo tedy mnoho znakových sad s různým významem kódů nad 127 (Viz kapitolu 2.3).

Výhody

Jednoduché kódování (pevná šířka znaku - 8bit). Malý objem dat.

Nevýhody

Malý rozsah znakové sady.

Existují stovky jednobytových sad (ISO, ČSN, další závislé na platformě, ...).

2.2 Vícebytové znakové sady - Unicode

S rozvojem Internetu je potřeba výměny dokumentů mezi různými platformami a různými národními jazyky. Řešením je vícebytová znaková sada (dále jen *Unicode*).

Do její přípravy se bohužel najednou pouštějí dvě různé organizace:

- ISO (International Organization for Standardization) od 1989,
  znaková sada ISO/IEC 10646, ve zkratce *UCS* (Universal Character Set).
- Unicode (Unicode Consortium) od 1990, znaková sada Unicode, ve zkratce *Unicode*.

Naštěstí obě organizace od roku 1991 spolupracují. V současnosti jsou číselné hodnoty (code points) jednotlivých znaků identické. Liší se kódovací schéma.

Vývoj Unicode

1991 - 16bitová znaková sada: *BMP* (Basic Multilingual Plane)

1996 - rozšíření na 32 bitů (používáno 21 bitů, tj. asi milion znaků)

Kódovací schéma

UTF-32: každý znak - 4 byty

UTF-16: každý znak v BMP - 2 byty, ostatní znaky - 4 byty

UTF-8: proměnná délka 1-6 bytů na znak (prakticky 1-4)

Další: UCS-2, UCS-4, UTF-16BE, UTF-16LE, UTF-32BE, UTF-32LE

UTF-8

Výhody

Znaky US-ASCII se kódují stejně => kompatibilita.

Pro akcentované znaky stačí 2 byty => relativní úspora dat.

Nevýhody

Znaky nemají stejnou délku => není možné skočit o určitý počet znaků.

2.3 Znakové sady pro češtinu

Pojmenování charsetů se řídí Internet Assigned Numbers Authority (IANA).

V závorce jsou uvedeny alternativní označení.

1. *ISO-8859-2* (ISO Latin 2, IBM912): Osmibitové kódování češtiny v UNIXových systémech. Specifikováno normou ISO-8859-2 (z roku 1987).
2. *Bratři Kameničtí* (KEYBCS2, CP895, MJK)
3. *Cork* (T1): Kódování Cork používá většina evropských TUG (národní TeX Users Groups) pro TeXovské mezinárodní písmo T1.
4. *CP852* (IBM852, PC Latin 2, PC L2): Osmibitové kódování češtiny v systému MS-DOS (Konzole v MS Windows). Toto kódování má všechny tisknutelné znaky sady ISO-8859-2 (ISO Latin 2), diakritická písmenka jsou však na jiných pozicích.
5. *KOI8-ČS* (KOI8ČS): Toto kódování bylo používáno na starých terminálech. Např. program T602 stále dovoluje jeho používání. Obsahuje "ch"jako zvláštní znak.
6. *Mac OS Central European* (MacCE, CE, Mac, apple-ce, x-mac-ce...): Tuto znakovou sadu používá lokalizovaný Mac OS (především počítače Apple Macintosh).
7. *Windows-1250* (CP1250, WinCS, WinEE): Téměř shodné s ISO Latin 2. Čtrnáct znaků je však na jiných pozicích. Písma s koncovkou CE.
8. *Unicode* (UTF-8)

2.4 Problémy textového formátu

Různé řídící znaky v různých operačních systémech
(např. konec řádku Unix - LF, Mac - CR, MS - CR+LF).

Různé znakové sady s různým rozsahem a různá kódovací schémata (charset).

Řešení

U dokumentů uvádět (např. v hlavičce HTML) použitý charset.

Používat software umožňující rozpoznání a konverzi.

Snaha o sjednocení => jednotný národní charset, jednotný světový Unicode.

3 Binární formát

Výhody

Vhodné pro okamžité zpracování (formát dat shodný s tvarem v operační paměti).

V případě poškození části dat těžko opravitelné.

Nevýhody

Možnost utajení formátu.

Nutnost použití specifického programu.

Složitější interpretace, viry, ...

4 Otevřené a uzavřené formáty

Otevřený formát

Jeho specifikace je volně dostupná.

Většinou lépe přenositelný (umí ho číst více programů ve více operačních systémech).

Je prostředkem pro výměnu informací, efektivní využití a zpracování dat.

Např.: ODF, prostý text, ...

Uzavřený formát

Umožňuje získat monopol pro jeho zpracování a zároveň silně omezuje možnosti využití uložených dat. Např. MS Office.

5 Konverze formátů

Často se dostaneme do situace, kdy požadujeme data ve specifickém formátu, který se liší od toho, který máme k dispozici. Např. spolupráce programů, z nichž každý pracuje s jiným formátem.

Konverzí formátu se rozumí v ideálním případě jeho změna bez ztráty, nebo nabytí informačního obsahu.

To zdaleka ne vždy platí. Viz kapitolu 5.3.

5.1 Rozpoznání (čtení)

Pro správnou konverzi je nutné znát formát souboru, který chceme převést.

Pro textový formát je podstatné v jakém je kódování a z jakého operačního systému.

V případě binárního souboru nám napoví přípona. Pokud ji soubor nemá musíme použít speciální program. Nejkrajnější nouzové řešení je pak zkoušení, který program soubor přečte + editace přípony.

5.2 Provedení konverze

5.2.1 Speciálním programem

Test mnoha programů pro změnu kódování

`http://vorisekd.wz.cz/test.htm`

Jsou i programy pro konverzi zpřístupněné on-line

`http://media-convert.com/`, `http://www.zamzar.com/`

5.2.2 Službami Open a Save (As) běžných programů

Např. pomocí OpenOffice.org lze pomocí *uložit jako* konvertovat DOC na ODT a opačně. V textovém editoru PSPad lze editačními nástroji měnit charset i typ zalamování řádků.

Tento postup je však vhodný pro konverzi velkého množství souborů až se současným použitím maker. Více např na `http://www.openoffice.cz/navody/`.

5.3 Problémy konverze

5.3.1 Různý rozsah znakových sad

U textových formátů se setkáváme s problémem vyjádření znaků, které nemají ve výstupním formátu ekvivalent. Řešením můžou být:

ESC sekvence a HTML kódy: Jeden nebo několik znaků následujících znak *ESC* (v HTML &) nejsou interpretovány jako ASCII kódy.
Např. Ä můžeme v HTML zapsat jako &*Auml*;.

Transliterace: Např. `ť` lze aproximovat pomocí `t` nebo `t'`.

5.3.2 Textové formátovací značky

... užívají pouze základní US-ASCII, takže zde problémy převážně nejsou.

5.3.3 Binární formát

Zde je především obtížná interpretace vstupního souboru a řešením může být snad jen zjednodušení požadavků na výstupní formát, nebo volba kvalitnějšího konverzního nástroje. V případě uzavřených formátů jsou problémy obecně častější a závažnější.

Závěr

Text tohoto dokumentu je samozřejmě jen úvodem do problematiky. Obsahuje však dost odborných pojmů, které čtenáři pomůžou při studování další literatury.

Věřím, že tento dokument poslouží i k řešení praktických problémů s kódováním a konverzí.

Literatura

[1] Rybička, J.: WWW stránky, [online], [cit. 2009-03-29], URL: <http://old.mendelu.cz/~rybicka/>

[2] Herout, P.: WWW stránky, [online], [cit. 2009-03-29], URL: <http://www-kiv.zcu.cz/~herout/>

[3] Wikipedia: WWW stránky, [online], [cit. 2009-03-29], URL: <http://en.wikipedia.org/>