BESEDILO

Zapis besedila

- Besedilo je sestavljeno iz znakov
  - kako so zapisani?
- **Nabor znakov** (*character repertoire*)
  - vsi znaki neke abecede, npr A, B, C, Č, a, b, c, 0, 1, 2, :, + ...
  - nabor ima lahko znake, ki izgledajo isto, so pa različni, npr. A (latinski A), A (grška alfa), A (A v cirilici)
- **Koda znaka** (*code position, code value, code point ...*)
  - določa število za vsak znak iz repertoarja
  - npr. ISO 10646 določa kode za znake "a", "!", "ä", and "‰" kot 97, 33, 228, and 8240
  - naboru vseh kod za repertoar lahko rečemo *coded character set (CCS)*
- **Kodiranje znakov** (*character encoding*)
  - določa kako se koda znaka dejansko zapiše v obliki bytov, npr niz a!ä‰ lahko zapišemo kot:
    - 61 21 C3 A4 E2 80 B0 (UTF-8) ali
    - 61 00 21 00 E4 00 30 20 (UTF-16)

ASCII kodiranje

- ASCII (American Standard Code for Information Interchange)
- Od 1970-ih let
- 7 bitna kodna tabela
  - 8. bit za popravljanje napak
  - 128 kod
  - 95 znakov
  - kontrolni znaki...
- Ameriška angleščina
- ISO 646 standard leta 1972
- Vsebuje tudi nacionalne variante

32		33	!	34	"	35	#
36	$	37	%	38	&	39	'
40	(	41	)	42	*	43	+
44	,	45	-	46	.	47	/
48	0	49	1	50	2	51	3
52	4	53	5	54	6	55	7
56	8	57	9	58	:	59	;
60	<	61	=	62	>	63	?
64	@	65	A	66	B	67	C
68	D	69	E	70	F	71	G
72	H	73	I	74	J	75	K
76	L	77	M	78	N	79	O
80	P	81	Q	82	R	83	S
84	T	85	U	86	V	87	W
88	X	89	Y	90	Z	91	[
92	\	93	]	94	^	95	_
96	`	97	a	98	b	99	c
100	d	101	e	102	f	103	g
104	h	105	i	106	j	107	k
108	l	109	m	110	n	111	o
112	p	113	q	114	r	115	s
116	t	117	u	118	v	119	w
120	x	121	y	122	z	123	{
124	|	125	}	126	~		

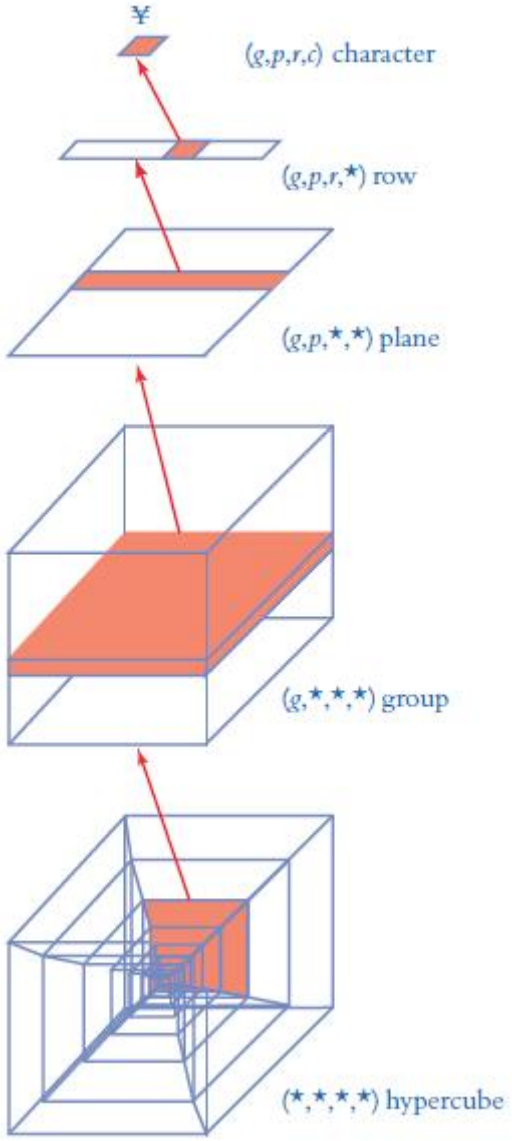
ISO Latin 8859 kodiranje

- 8 bitne kodne tabele
- Spodnja polovica ASCII
- Zgornja polovica glede na regijo
  - 10 delov
  - ISO 8859-1: ISO Latin1, Zahodna Evropa
  - ISO 8859-2: ISO Latin2, Vzhodna Evropa
  - ISO 8859-5: ISO Latin5, Cirilica
  - ISO 8859-8: ISO Latin8, Hebrejščina
- Še vedno ne omogoča uporabe več jezikov hkrati

160		161	¡	162	¢	163	£
164	¤	165	¥	166	¦	167	§
168	¨	169	©	170	ª	171	«
172	¬	173	-	174	®	175	¯
176	°	177	±	178	²	179	³
180	´	181	µ	182	¶	183	·
184	¸	185	¹	186	º	187	»
188	¼	189	½	190	¾	191	¿
192	À	193	Á	194	Â	195	Ã
196	Ä	197	Å	198	Æ	199	Ç
200	È	201	É	202	Ê	203	Ë
204	Ì	205	Í	206	Î	207	Ï
208	Ð	209	Ñ	210	Ò	211	Ó
212	Ô	213	Õ	214	Ö	215	×
216	Ø	217	Ù	218	Ú	219	Û
220	Ü	221	Ý	222	Þ	223	ß
224	à	225	á	226	â	227	ã
228	ä	229	å	230	æ	231	ç
232	è	233	é	234	ê	235	ë
236	ì	237	í	238	î	239	ï
240	ð	241	ñ	242	ò	243	ó
244	ô	245	õ	246	ö	247	÷
248	ø	249	ù	250	ú	251	û
252	ü	253	ý	254	þ	255	ÿ

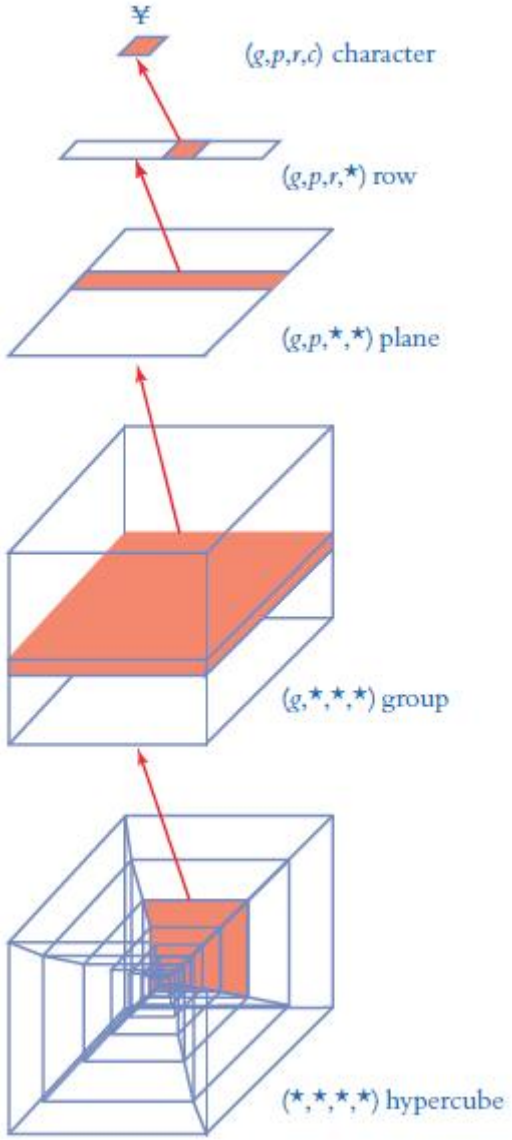
ISO 10646

- UCS - *Universal Character Set*
- 32 bitov
  - 4-dimenzionalna kocka
  - 256 skupin
  - 256 ravnin
  - 256 vrst
  - 256 znakov (8-bitna množica znakov)
  - (g,p,r,c)
- (0,0,0,*) = ISO Latin1



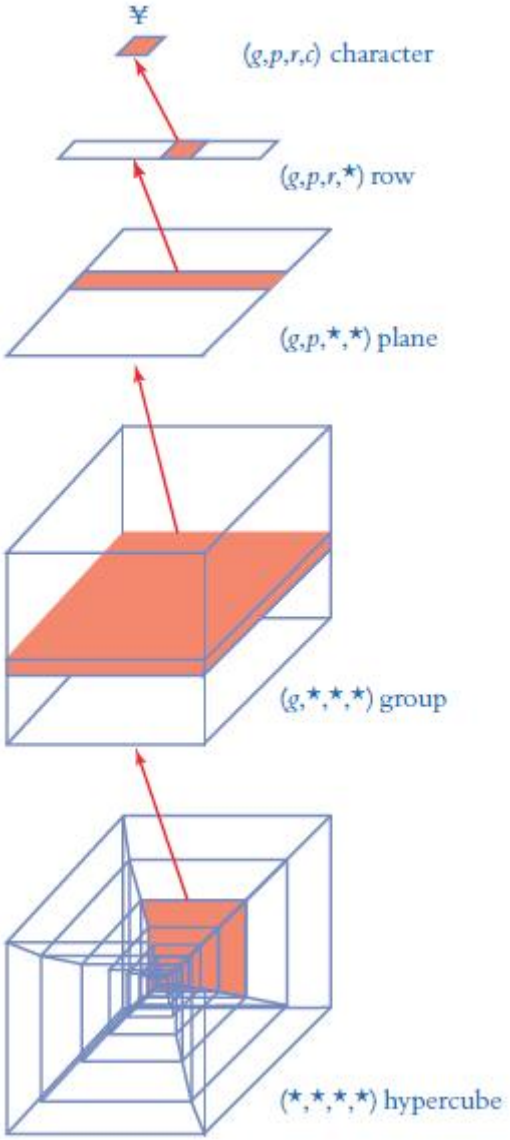
Unicode

- Industrijski standard, združljiv z UCS
  - postavlja dodatne omejitve glede implementacije
- Trenutno kode med od 0..10FFFF
  - 0-FFFF Basic Multilingual Plane (BMP)
    - Latinica, Cirilica, Grščina, Sudanščina, Tibetanščina, vzhodno-azijske pisave ...
  - 10000-1FFFF Supplementary Multilingual Plane
    - Gotika, Staroperzijščina, bizantinski glasbeni simboli, emotikoni, alkemijski simboli ...
  - 20000-2FFFF Supplementary Ideographic plane
    - dodatni znaki za vzhodnoazijske jezike



UCS/Unicode Kodiranja

- UCS-4 ~= UTF-32
  - štiri byte za vsako kodo znaka (potratno)
- UTF-16
  - BMP predstavi z dvema bytoma, višje ravnine s štirimi byti (prva dva sta rezervirana koda iz BMP)
  - potrebno še definirati ali je big endian ali little endian zapis
- UTF-8
  - ASCII znake 00-7F predstavi z enim bytom
  - ostale z dvema do štirimi byti med 80-FF
- UTF-7
  - predstavi znake z enim ali več byti med 00-7F



Unicode Byte Order Mark

- Za določanje vrstnega reda bytov (big/little endian) pri UTF-16 in UTF-32, lahko na začetku datoteka vsebuje t.i. BOM, preko katerega bralec lahko prepozna vrstni red
  - zapis posebnega znaka s kodo U+FEFF
  - UTF-8 datoteke lahko tudi vsebujejo BOM, vendar bo vedno enako zapisan, lahko pa z njim prepoznamo, da je to UTF-8 datoteka

Bytes	Encoding Form
00 00 FE FF	UTF-32, big-endian
FF FE 00 00	UTF-32, little-endian
FE FF	UTF-16, big-endian
FF FE	UTF-16, little-endian
EF BB BF	UTF-8

Besedilo in oblika

- Najpogostejši zapis oblikovanega besedila je z **označevalnimi jeziki**
  - HTML, LaTeX, Office Open XML ...
- Oznake **posamično** vizualno oblikujejo izpis (pisavo, odstavek,...)
  - oznake določajo **logične elemente** dokumenta (naslovi, seznami, tabele, ipd.)
  - vsi takšni elementi v besedilu so enako formatirani
  - konsistentna postavitev v celem dokumentu
  - hitro spreminjanje oblike celotnega besedila
- Posebej je določena **oblika** (CSS ipd.) elementov
- Za **analizo** besedila potrebujemo **pregledovalnik** (parser)
  - oznake lahko pomagajo pri ugotavljanju pomembnosti
    - kaj je naslov ...

HTML

```
<h1>
This is a title
</h1>
<p>
  A <b>character encoding</b> system consists
  of a code.
</p>
<p>
  Other terms like <b>character encoding</b>.
</p>
```

CSS

```
. h1 {
  color: #00FF00;
  font-family:arial;
  font-size: 14pt;
}
```

REFERENCE

- A tutorial on character code issues
- wiki: Unicode
- UTF FAQ: BOM