

---

# **Zapis teksta u računaru**

---

---

- **Tekst (ili dokument)** je

- "informacija namenjena ljudskom sporazumevanju koja može biti prikazana u dvodimenzionalnom obliku... Tekst se sastoji od grafičkih elemenata kao što su karakteri, geometrijski ili fotografski elementi ili njihove kombinacije, koji čine sadržaj dokumenta." (ISO-definicija)

---

# TEKST JE ... (NIZ KARAKTERA)

- Iako obično tekst zamišljamo kao dvodimenzioni objekat, u računarima se tekst predstavlja kao jednodimenzioni (linearni) niz karaktera.
- Potrebno je, dakle, uvesti specijalne karaktere koji označavaju prelazak u novi red, tabulator, kraj teksta i slično

---

# ZAPIS KARAKTERA U RAČUNARU

- Računari su zasnovani na binarnoj aritmetici
- Cele brojeve je moguće predstaviti u binarnom sistemu
- Osnovna ideja je svakom karakteru pridružiti određeni ceo broj na unapred dogovoreni način
- Ove brojeve zovemo ***kodovima karaktera*** (character codes)

---

# GRAFIČKA REPREZENTACIJA KARAKTERA

- Glif – grafička reprezentacija karaktera
- Font – skup glifova

---

# KOLIKO KARAKTERA ŽELIMO DA PREDSTAVIMO U RAČUNARIMA?

- Tokom razvoja računarstva broj karaktera je postajao sve veći
- Pošto je u početku razvoja englesko govorno područje bilo dominantno osnovno je bilo predstaviti sledeće karaktere:

---

# ENGLESKO GOVORNO PODRUČJE

- Velika slova engleskog alfabetu : A,B,...,Z
- Mala slova engleskog alfabetu : a,b,...,z
- Cifre : 0,1,...,9
- Interpunkcijske znake : .,:;'+\*-\_ i slično
- Specijalne znake : kraj reda, tabulator i slično

---

## STANDARDNI KARAKTERSKI KODOVI

- Šezdesetih godina su se pojavile tabele standardnih karakterskih kodova dovoljne za zapis pomenutih karaktera
- Najpoznatiji su
  - EBCDIC – IBM-ov standard, korišćen uglavnom na mainframe računarima, pogodan za bušene kartice
  - ASCII – Standard iz koga se razvila većina današnjih standarda



---

# ASCII

- **ASCII (*American Standard Code for Information Interchange*)**
- Uspostavljen od strane organizacije ANSI (American National Standard Institute)
- ASCII sedmobitan (broj karaktera je 128)

# ASCII TABELA

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

---

## PRIMERI

- Karakter **A** se zapisuje kao  $(41)_{16}$  tj.  $0x41$  što je  $(65)_{10}$  tj.  $(1000001)_2$
- Razmak se zapisuje kao  $(20)_{16}$  što je  $(32)_{10}$  tj.  $(0100000)_2$

Primer 1. Zapišite cifru **3** u ASCII kodu

Primer 2. Zapišite tekst *Fakultet* u ASCII kodu

---

## Rešenje primera 1:

- Cifra: *3*
- Heksadekadno ASCII: *33*
- Dekadno ASCII: *51*
- Binarno ASCII: *0110011*

---

Rešenje primera 2:

■ Tekst: *Fakultet*

■ Heksadekadno ASCII:

*46 61 6B 75 6C 74 65 74*

■ Dekadno ASCII:

*70 97 107 117 108 116 101 116*

■ Binarno ASCII:

*1000110 1100001 1101011 1110101  
1101100 1110100 1100101 1110100*

---

# OZNAKA ZA KRAJ REDA

- Oznaka za kraj reda se ne zapisuje isto u svim operativnim sistemima
- Pod DOS/Windows oznaka za kraj reda se zapisuje sa dva karaktera (**CR LF - Carriage return Line feed**), 0xD 0xA tj. 13 10 – istorijski razlozi (stari štampači)
- Unix koristi samo karakter **LF** tj. 0xA

---

# Grčki alfabet, ćirilčni tekstovi,...

- Razvojem računarstva se javlja potreba kodiranja tekstova i na drugim jezicima
- Kroz istoriju su postojala mnoga rešenja, od kojih su se neka zadržala, a neka su nestala

---

# KODNE STRANE

- Pod *kodnom stranom* (*Code page*) tj. *skupom karaktera* (*Character set, charset*) podrazumevamo uređenu listu karaktera predstavljenih svojim karakterskim kodovima



---

# KODNE STRANE

- Podaci se u računarima obično zapisuju bajt po bajt
- ASCII je sedmobitni standard
- ASCII karakteri se zapisuju tako što se u svakom bajtu bit najveće težine postavi na 0
- To ostavlja prostor za novih 128 karaktera čiji binarni zapis počinje sa 1

---

## KODNE STRANE

- Ovaj prostor se može popuniti na razne načine
  - Rešenje nije univerzalno, jer svakako na svetu postoji više od 256 različitih karaktera
  - Postavljeni su razni standardi dopunjavanja ovih 128 karaktera
  - Svim ovim kodnim stranama je zajedničko prvih 128 karaktera i oni se poklapaju sa ASCII
-

---

## KODNE STRANE

- Ovako napravljene kodne strane obično omogućuju kodiranje tekstova na više srodnih jezika (obično i geografski bliskih)
- Nama su uglavnom važne kodne strane napravljene za centralno-evropske (Central European) latinice, kao i ćirilične kodne strane

---

# NAJČEŠĆE KORIŠĆENE KODNE STRANE KOD NAS

- **ISO 8859-2 (Latin2)**
  - **ISO 8859-5 (Ћирилична)**
  - **Windows 1250**
  - **Windows 1251 (Ћирилична)**
- 
- Prve dve su delo međunarodne organizacije za standardizaciju (ISO - International Standard Organization), dok su naredne dve Microsoft-ovi standardi

---

# LATIN 1

- Poželjno je poznavati i osnovnu kodnu stranu **ISO 8859-1 (Latin1)** jer je veoma često postavljena kao podrazumevana kodna strana. Ona se koristi za zapis tekstova na zapadno evropskim jezicima (Western European)

# ISO 8859-1 (LATIN1)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8																
9																
A	NBSP	ı	ç	£	¤	¥	ı	§	"	©	ª	«	¬	SHY	®	ˆ
B	°	±	²	³	´	µ	¶	·	,	¹	º	»	¼	½	¾	¿
C	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
E	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
F	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	

# ISO 8859-2

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8																
9																
A	NBSP	Ą	˘	Ł	▣	Ł	Ś	§	˝	Š	Ş	Ť	Ž	SHY	Ž	Ž
B	°	ą	˙	ł	◻	ł	ś	¶	˚	š	ş	ť	ž	”	ž	ž
C	Ŕ	Á	Â	Ã	Ä	É	Ć	Ç	Č	É	Ę	Ë	Ě	Í	Î	Ď
D	Đ	Ń	Ň	Ó	Ô	Õ	Ö	×	Ř	Ů	Ú	Û	Ü	Ý	Ť	ß
E	ŕ	á	â	ã	ä	é	ć	ç	č	é	ę	ë	ě	í	î	ď
F	đ	ń	ň	ó	ô	õ	ö	÷	ř	ů	ú	û	ü	ý	ț	

# WINDOWS 1250

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	€		,		„	…	†	‡		‰	Š	<	Ś	Ť	Ž	Ž
9		,	,	“	”	•	–	–		™	š	>	ś	ť	ž	ž
A		˘	˘	Ł	ł	Ą	ą	§	¨	©	Ş	«	¬		®	Ž
B	°	±	˙	ł	´	μ	¶	·	˘	ą	ş	»	Ł	˝	ł	ž
C	Ř	Á	Â	Ă	Ä	Ĺ	Ć	Ç	Č	É	Ę	Ě	Ě	Í	Î	Ď
D	Ð	Ń	Ñ	Ó	Ô	Õ	Ö	×	Ř	Û	Ú	Û	Ü	Ý	Ť	ß
E	ř	á	â	ă	ä	ĺ	ć	ç	č	é	ę	ě	ě	í	î	ď
F	đ	ń	ñ	ó	ô	õ	ö	÷	ř	û	ú	û	ü	ý	ț	



# ISO-8859-5

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8																
9																
A	NBSP	Ё	Ђ	Ѓ	Є	Ѕ	І	Ї	Ј	Љ	Њ	Ћ	Ќ	SHY	Ў	Џ
B	А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
C	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
D	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
E	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я
F	№	ё	ђ	ѓ	є	ѕ	і	ї	ј	љ	њ	ќ	ќ	§	ў	

---

## VIŠEBAJTNI KARAKTERSKI KODOVI

- Iako navedene kodne strane omogućuju kodiranje tekstova koji nisu na engleskom jeziku nije moguće npr. u istom tekstu mešati ćirilicu i našu latinicu.
- Azijskim jezicima nije dovoljno 256 mesta za zapis svih karaktera.
- Zbog toga se uvode višebajtni karakterski kodovi

## UCS, ISO 10646, UNICODE

- Kasnih osamdesetih, dve velike organizacije su pokušale standardizaciju tzv. Univerzalnog skupa karaktera (Universal Character Set - UCS)
- To su bili ISO, kroz standard 10646 i projekat UNICODE organizovan i finansiran uglavnom od strane američkih firmi koje su se bavile proizvodnjom višejezičkog softvera (Apple, Sun Microsystems, Microsoft,...).

---

# ISO 10646

- ISO 10646 je zamišljen kao 4 bajtni standard. Pri tome se prvih 65536 ( $2^{16}$ ) karaktera koriste kao osnovni višejezični skup karaktera dok je ostali prostor ostavljen kao proširenje za drevne jezike, celokupnu naučnu notaciju i slično.

# UNICODE

- Početna verzija Unicode standarda podrazumevala je dvobajtni kod za svaki karakter (heksadekadno,  $(0000)_{16}$ - $(FFFF)_{16}$ )
- Vremenom se shvatilo da  $2^{16}=65536$  karaktera nije dovoljno za zapis svih karaktera koji postoje, pa je odlučeno da se skup kodova proširi na  $2^{20}=1048576$  (heksadekadno,  $(000000)_{16}$ - $(10FFFF)_{16}$ )
- Prošireni skup kodova podeljen je na 16 takozvanih ravni pri čemu svaka ravan sadrži  $2^{16}=65536$  karaktera

---

# UNICODE

- Prvih  $2^{16}=65536$  karaktera kodiranih u opsegu  $(0000)_{16}-(FFFF)_{16}$  čini takozvanu osnovnu višejezičku ravan, u koju spada većina danas korišćenih karaktera uključujući čak i često korišćene CJK karaktere:
- Vremenom su se ISO 10646 i Unicode združili i danas postoji izuzetno preklapanje između ova dva standarda

## UNICODE

- Raspored određenih grupa karaktera u osnovnoj višejezičkoj ravni:

0200-007E – ASCII printable

00A0 – 00FF Latin-1

0100 – 017F – Latin Extended A

(osnovno proširenje latinice, sadrži sve naše dijakritike)

0180-077F – Latin Extended B

...

0370-03FF Ćirilica

## UCS-2

- Unicode standard u suštini predstavlja veliku tabelu koja svakom karakteru dodeljuje broj.
- Standardi koji opisuju kako se niske karaktera onda prevode u nizove bajtova se dodatno definišu
- ISO definiše UCS-2 standard koji jednostavno svaki UNICODE karakter osnovne višejezičke ravni prevodi u odgovarajuća dva bajta



---

## UTF

- Latinični tekstovi kodirani preko UCS-2 standarda sadrže veliki broj nula, koje obično u operativnim sistemima poput UNIX-a i u programskom jeziku C imaju specijalno značenje.
- Iz istog razloga softver koji je razvijen za rad sa dokumentima u ASCII formatu ne može da radi bez izmena nad dokumentima kodiranim preko UCS-2 standarda

---

# UTF

- *Unicode transformation format (UTF)* algoritam svakom UNICODE karakteru dodeljuje određeni niz bajtova čija dužina varira od 1 do najviše 3.
- UTF je ASCII kompatibilan, što znači da se ASCII karakteri zapisuju pomoću jednog bajta, na standardni način.

# UTF-8

raspon	binarno zapisan Unicode kôd	binarno zapisan UTF-8 kôd
0000-007F	00000000 0xxxxxxx	0xxxxxxx
0080-07FF	00000yyy yyxxxxxx	110yyyyy 10xxxxxx
0800-FFFF	zzzzyyyy yyxxxxxx	1110zzzz 10yyyyyy 10xxxxxx

---

Zapisati reči MATF u kodnim stranama ISO 8859-2, Windows 1250, Windows 1251.

- Reč **MATF** se zapisuje isto u kodnim stranama ISO 8859-2, Windows 1250 i Windows 1251 zato što su njeni karakteri zapravo ASCII karakteri (jer svim ovim kodnim stranama zajedničko je prvih 128 karaktera i oni se poklapaju sa ASCII kodovima).
- Dakle, reč MATF se u ovim kodnim stranama zapisuje preko 4 bajta i to

$(4D)_{16}, (41)_{16}, (54)_{16}, (46)_{16}$

- To je isto što i

$(77)_{10}, (65)_{10}, (84)_{10}, (70)_{10}$

- Tj.

$(01001101)_2, (01000001)_2, (01010100)_2,$

$(01000110)_2.$

---

---

## Zapisati reči lišće u kodnim stranama ISO 8859-2, Windows 1250, Windows 1251

- Reč *lišće* sadrži u sebi karaktere š i ć koji nisu ASCII karakteri pa se različito kodiraju u svakoj od kodnih strana.
  - U kodnoj strani ISO 8859-2 odgovarajući kod je  $(6c)_{16}$ ,  $(69)_{16}$ ,  $(b9)_{16}$ ,  $(e6)_{16}$ ,  $(65)_{16}$
  - To je isto što i  $(108)_{10}$ ,  $(105)_{10}$ ,  $(185)_{10}$ ,  $(230)_{10}$ ,  $(101)_{10}$
  - Tj.  $(01101100)_2$ ,  $(01101001)_2$ ,  $(10111001)_2$ ,  $(11100110)_2$ ,  $(01100101)_2$ .
  - U kodnoj strani Windows 1250 karakteri š i ć se kodiraju sa  $(9A)_{16}$ ,  $(E6)_{16}$
  - U kodnoj strani Windows 1251 karakteri š i ć se ne mogu kodirati.
-