

Αναπαράσταση Μη Αριθμητικών Δεδομένων

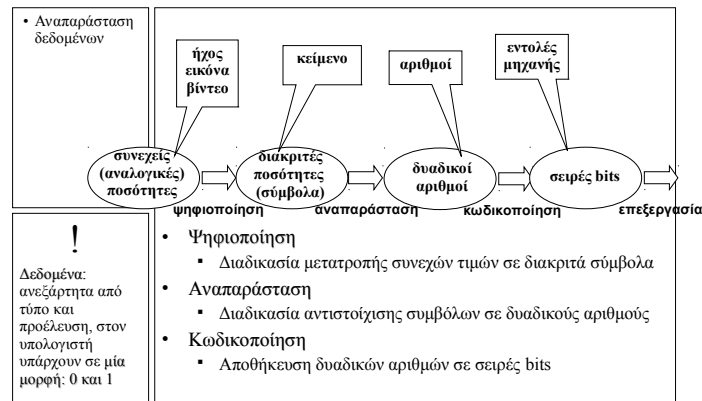
(κείμενο, ήχος και εικόνα στον υπολογιστή)

<http://mixstef.github.io/courses/csintro/>



Μ.Στεφανιδάκης

Αναπαράσταση δεδομένων



Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

2

Η ερμηνεία της αναπαράστασης

• Αναπαράσταση δεδομένων

- Κάπου στη μνήμη του υπολογιστή...
 - Βρίσκεται αποθηκευμένη η σειρά bits 0100110111010001
- Πόσα σύμβολα αναπαριστά;
 - Πόσα bits ανά σύμβολο;
- Ποιος ο τύπος των δεδομένων;
- Ποια συγκεκριμένη ποσότητα συμβολίζει;
- Πώς θα το χειριστεί ο υπολογιστής;

! Στα ερωτήματα αυτά μπορεί να απαντήσει μόνο ο προγραμματιστής της εφαρμογής που χειρίζεται τα δεδομένα!

Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

3

Αναπαράσταση με δυαδικούς αριθμούς

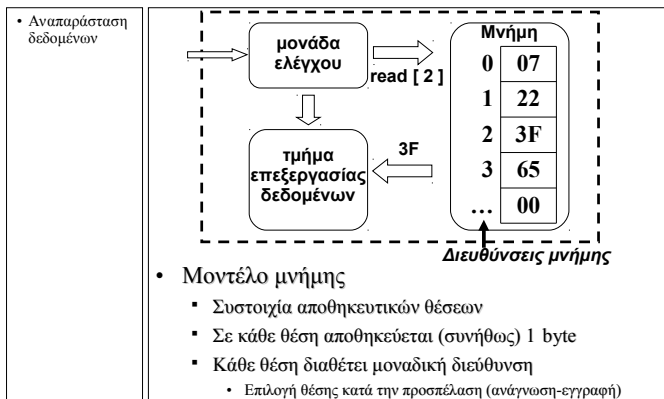
• Αναπαράσταση δεδομένων

- Σειρά n bits
 - Δυαδικός αριθμός με n bits ($n \geq 1$) μπορεί να αναπαράσχει 2^n διαφορετικά σύμβολα
- Μη αριθμητικά δεδομένα
 - Κείμενο, εντολές μηχανής, ήχος, εικόνα...
 - Σύνολο διαφορετικών αντικειμένων (συμβόλων)
 - Αντιστοίχιση κάθε συμβόλου σε μοναδικό δυαδικό αριθμό (code point)
 - “Αναπαράσταση”
 - Η ακριβής αντιστοίχιση συνήθως ορίζεται σε ένα πρότυπο (standard)

Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

4

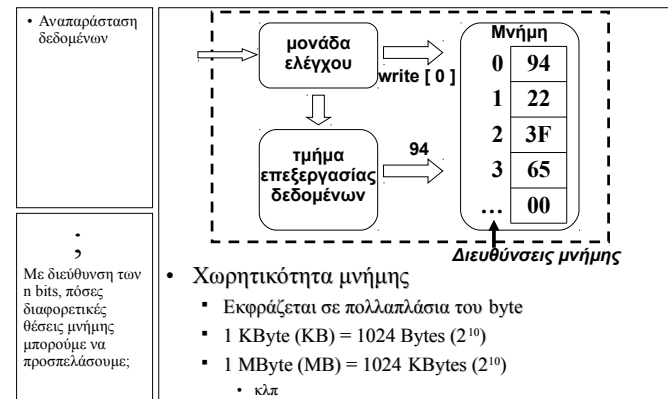
Η επικοινωνία με τη μνήμη



Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

5

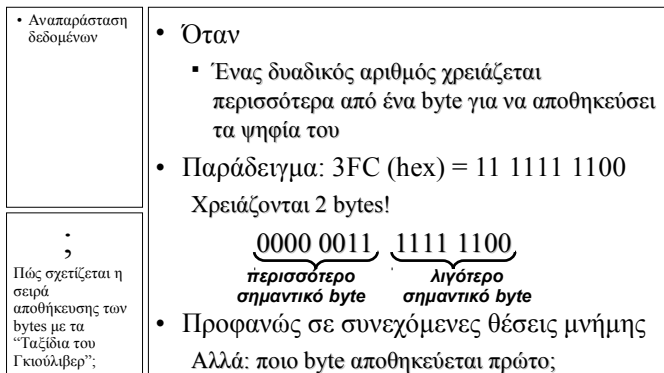
Η επικοινωνία με τη μνήμη



Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

6

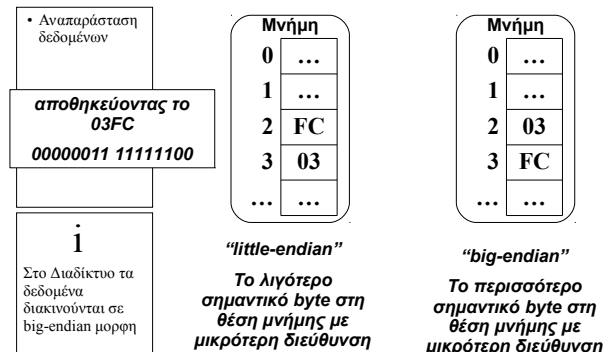
Θέματα αποθήκευσης δυαδικών αριθμών



Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

7

Θέματα αποθήκευσης δυαδικών αριθμών



Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

8

Αρχικές αναπαράστασεις κειμένου

- Αναπαράσταση δεδομένων
- Κείμενο

- Οι πρώτες αναπαράστασεις κειμένου
 - Στον υπολογιστή
 - 6-7 bits ανά χαρακτήρα
 - Πόσοι διαφορετικοί χαρακτήρες;
- Μη εκτυπώσιμοι χαρακτήρες
 - Χαρακτήρες ελέγχου
 - Ιδιαίτερα χρήσιμοι για τις συσκευές εξόδου της εποχής (εκτυπωτές, τηλετυπα...)
 - Νέα γραμμή (LINE FEED – LF)
 - Επιστροφή κεφαλής εκτύπωσης (CARRIAGE RETURN – CR)
 - Καμπανάκι (BELL) κλπ

Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

9

Κώδικας ASCII

- Αναπαράσταση δεδομένων
- Κείμενο

- Βασικό αρχικό πρότυπο αναπαράστασης κειμένου
 - 7 bits ανά χαρακτήρα

STANDARD ASCII ΚΩΔΙΚΑΣ					
hex	char	hex	char	hex	char
20		40	@	60	
21	!	41	A	61	a
22	"	42	B	62	b
23	#	43	C	63	c
24	\$	44	D	64	d
25	%	45	E	65	e
26	&	46	F	66	f
27	'	47	G	67	g
28	(48	H	68	h
29)	49	I	69	i
2A	*	4A	J	6A	j
2B	+	4B	K	6B	k
2C	,	4C	L	6C	l
2D	-	4D	M	6D	m
2E	.	4E	N	6E	n

i
ASCII: American
Standard Code for
Information
Interchange

Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

10

Κείμενο σε κώδικα ASCII

- Αναπαράσταση δεδομένων
- Κείμενο

;
Με 7 bits ανά χαρακτήρα και χρήση bytes, 1 bit μένει αχρησιμοποίητο. Πόσοι επιπλέον χαρακτήρες με το bit αυτό;

- 7 bits ανά χαρακτήρα
 - 128 χαρακτήρες
 - Αναπαράσταση με τους αριθμούς 0...127
- Κανονικοί χαρακτήρες (εκτυπώσιμοι)
 - 32...64, 91...96, 123...126 = σημεία στίξης κ.ά. (32 = SPACE!)
 - 65...90 = κεφαλαία λατινικά (A-Z)
 - 97...122 = πεζά λατινικά (a-z)
- Χαρακτήρες ελέγχου (μη εκτυπώσιμοι)
 - 0...31, 127 – επιζούν τα: 9 (TAB), 13/10 (CR/LF, σήμανση “νέας γραμμής”)

Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

11

Κείμενο σε κώδικα ASCII

- Αναπαράσταση δεδομένων
- Κείμενο

- Παράδειγμα

H	a	v	e		a		n	i	c	e		d	a	y	!
72	97	118	101	32	97	32	110	105	99	101	32	100	97	121	33

!
Εφόσον η κωδικοποίηση είναι με 1 byte ανά χαρακτήρα, δεν τίθεται θέμα “little-” ή “big-endian”

- Γλώσσες προγραμματισμού
 - Συμβολοσειρά (string)
 - Σε γλώσσες όπως η C, το 0 (αριθμητικό) συμβολίζει το τέλος της συμβολοσειράς
 - Ο υπολογιστής μπορεί να κάνει πράξεις (π.χ. σύγκριση) με τη συμβολοσειρά

Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

12

Επεκτάσεις κώδικα ASCII

- Αναπαράσταση δεδομένων
- Κείμενο

!

Χρησιμοποιώντας τον ISO-8859-1 δεν είναι δυνατή η αναπαράσταση των ελληνικών!

- Χρήση του 1 επιπλέον bit του byte
 - 128 + 128 χαρακτήρες, αριθμοί 0...255
 - 0...127 αντιστοιχούν στον αρχικό ASCII
 - 127...255: επεκταμένα αλφάβητα
- Επέκταση αλφαβήτων (πρότυπα)
 - Χαρακτήρες που δεν υπάρχουν στον ASCII
 - Διαφορετικά ανά γλώσσα! Π.χ.:
 - ISO-8859-1: Δυτική Ευρώπη (Å, Ñ, Æ, ø κλπ)
 - ISO-8859-7: Νέα Ελληνικά
 - ...και πολλά άλλα πρότυπα για τις υπόλοιπες γλώσσες
 - Επίσης: μη πρότυπες λύσεις
 - Για Windows, Mac ..

Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

13

Κώδικας ISO-8859-7

- Αναπαράσταση δεδομένων
- Κείμενο

	x0	x1	x2	x3	x4	x5	x6	x7	x8	x9	xA	xB	xC	xD	xE	xF
0x	unused															
1x	unused															
2x	SP	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{		}	~	
8x	unused															
9x	unused															
Ax	NBSP	·	´	ε	€	Δ	Ⓜ	!	5	ˆ	Ⓜ	.	~	SHY	—	
Bx	•	±	²	³	´	ˆ	˜	Α	Β	Γ	Δ	Ε	Ζ	Η	Θ	Ι
Cx	Ϊ	Α	Β	Γ	Δ	Ε	Ζ	Η	Θ	Ι	Κ	Λ	Μ	Ν	Ξ	Ο
Dx	Π	Ρ	Σ	Τ	Υ	Φ	Χ	Ψ	Ω	Ϊ	Ϋ	ά	έ	ή	ί	
Ex	ύ	α	β	γ	δ	ε	ζ	η	θ	ι	κ	λ	μ	ν	ξ	ο
Fx	π	ρ	ς	σ	τ	υ	φ	χ	ψ	ω	ϊ	ϋ	ό	ύ	ώ	

[Wikipedia]

Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

14

Κείμενο σε κώδικα ISO-8859-7

- Αναπαράσταση δεδομένων
- Κείμενο

!

Οι αναπαράστασεις αλφαβήτων με 1 byte ανά χαρακτήρα τείνουν να καταρτηθούν!

- Παράδειγμα

Γ	ε	ι	α		σ	ο	υ	!
195	229	233	225	32	243	239	245	33
- Επέκταση κώδικα ASCII
 - 0...127 όπως στον ASCII
 - 128...159 πρόσθετοι χαρακτήρες ελέγχου
 - 160...255 ελληνικά και σχετικά σύμβολα

Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

15

Πρότυπο Unicode

- Αναπαράσταση δεδομένων
- Κείμενο

!

Με περισσότερα από 1 bytes ανά χαρακτήρα τίθεται θέμα σειράς αποθήκευσης των bytes!

- Για την αναπαράσταση όλων των αλφαβήτων!
 - Έχουν οριστεί σχεδόν 100.000 χαρακτήρες
 - Καλύπτει ιδεογράμματα, φωνητικές αναπαραστάσεις κλπ
 - Θα μπορούσε να καλύψει πάνω από 1 εκ. χαρακτήρες! (0 ... 10FFFF)
 - Κάθε χαρακτήρας αναπαρίσταται με περισσότερα από ένα bytes
 - Συνήθεις κωδικοποιήσεις: UCS-2 (ή UTF-16) και UTF-8
 - Το πρότυπο Unicode περιέχει επίσης
 - πληροφορία ισοδύναμων ή παρόμοιων χαρακτήρων
 - οδηγίες συνδυασμών τόνων/διακριτικών και γραμμάτων

Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

16

Ελληνικά και Unicode

- Αναπαράσταση δεδομένων
- Κείμενο

Greek and Coptic 03FF

	037	038	039	03A	03B	03C	03D	03E	03F
0		ι	Π	ϋ	π	β	η	ζ	
1		A	P	α	ρ	θ	λ	ρ	
2		B		β	ς	Υ	ϣ	ς	
3		Γ	Σ	γ	σ	Υ	ϣ	j	
4		Δ	T	δ	τ	Υ	ϣ	Θ	
5		E	Y	ε	υ	φ	φ	€	
6		A	Z	Φ	ζ	φ	ω	h	ε

Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

17

Κείμενο σε Unicode

- Αναπαράσταση δεδομένων
- Κείμενο

Παράδειγμα

Γ	ε	ι	α		σ	ο	υ	!
915	949	953	945	32	963	959	965	33
0393	03B5	03B9	03B1	0020	03C3	03BF	03C5	0021

δεκαεξαδικό

Κωδικοποίηση UCS-2 (big-endian)

03	93	03	B5	03	B9	03	B1	00	20	03	C3	03	BF	03	C5	00	21
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

Κωδικοποίηση UCS-2 (little-endian)

93	03	B5	03	B9	03	B1	03	20	00	C3	03	BF	03	C5	03	21	00
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

18

Unicode σε κωδικοποίηση UTF-8

- Αναπαράσταση δεδομένων
- Κείμενο

Αναπαράσταση μεταβλητού μήκους

Unicode	Κωδικοποίηση UTF-8
00...7F	0xxxxxxx
80...7FF	110xxxxx 10xxxxxx
800...FFFF	1110xxxx 10xxxxxx 10xxxxxx
10000...10FFFF	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx

!

Η κωδικοποίηση UTF-8 τείνει να επικρατήσει σε όλα τα προγράμματα που χειρίζονται κείμενα Unicode!

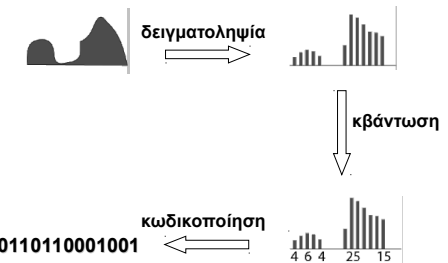
- Το βασικό λατινικό αλφάβητο (ASCII) χρησιμοποιεί 1 byte ανά χαρακτήρα
 - Προς τα πίσω συμβατότητα
- Τα ελληνικά, 2 bytes
 - Ποια η κωδικοποίηση κατά UTF-8 του τελευταίου παραδείγματος;

Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

19

Ήχος: Ψηφιοποίηση και Αποθήκευση

- Αναπαράσταση δεδομένων
- Κείμενο
- Ήχος



Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

20

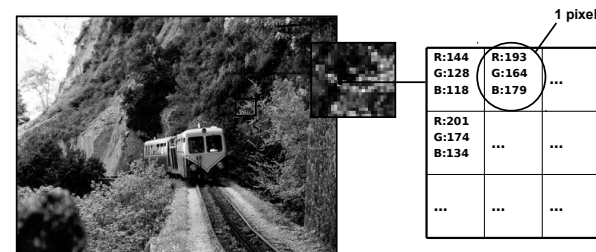
Εικόνα: από τον αναλογικό στον ψηφιακό κόσμο



Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

21

Παράδειγμα: απλή αναπαράσταση pixels με 16,7 εκ. χρώματα



- 3 bytes/pixel (24bits): R(ed) G(reen) B(lue)
 - 256 στάθμες ανά συνιστώσα χρώματος
 - $256 \times 256 \times 256 = 16.777.216$ χρώματα
 - εικόνες με μεγαλύτερο βάθος χρώματος
 - 32 έως 48 bits

Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

22

Εναλλακτικά: διανυσματικά γραφικά

- Αναπαράσταση δεδομένων
 - Κείμενο
 - Ήχος
 - Εικόνα
- Περιγραφή σχημάτων
 - Ως σύνολο ευθύγραμμων και καμπύλων τμημάτων
 - Με συντεταγμένες
 - Εύρεση σημείων μέσω μαθηματικού τύπου
 - Εύκολη αλλαγή μεγέθους γραφικών
 - Χωρίς παραμόρφωση των σχημάτων

Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

23

Αναπαράσταση βίντεο

- Αναπαράσταση δεδομένων
 - Κείμενο
 - Ήχος
 - Εικόνα
 - Βίντεο
- “Κινούμενη εικόνα” (καρέ)
 - όπως αναπαριστούμε τις απλές εικόνες
 - αλλά: με χρήση συμπίεσης
 - Για μείωση όγκου δεδομένων
 - Γειτονικά καρέ έχουν πολλές ομοιότητες

Εισαγωγή στην Επιστήμη των Υπολογιστών – “Αναπαράσταση Μη Αριθμητικών Δεδομένων”

24

Κωδικοποίηση εντολών μηχανής

- Αναπαράσταση δεδομένων
- Κείμενο
- Ήχος
- Εικόνα
- Βίντεο
- Εντολές Μηχανής

opcode	operand1	operand2	..	operandN
--------	----------	----------	----	----------

Περιγράφει το είδος της πράξης που θα εκτελεστεί

Περιγράφουν την προέλευση των δεδομένων εισόδου (αριθμό καταχωρητή, διεύθυνση μνήμης κλπ) και τον προορισμό των δεδομένων εξόδου (αποτέλεσμα πράξης)

Το είδος της πράξης προσδιορίζει τον τύπο, την προέλευση και τον αριθμό των δεδομένων που συμμετέχουν στην πράξη !