



Κεφ. 3: **Αναπαράσταση** **αλφαριθμητικών** **χαρακτήρων και** **αριθμών**

3.1.2 Κώδικας χαρακτήρων UNICODE



Ο κώδικας Unicode είναι ένα παγκόσμιο σύνολο χαρακτήρων (universal character set), δηλ. ένα διεθνές πρότυπο που παρέχει τη δυνατότητα κωδικοποίησης όλων των χαρακτήρων των σημαντικότερων γλωσσών του κόσμου.

Αποσκοπεί να είναι ένα υπεрсύνολο όλων των άλλων κωδικοποιήσεων συνόλων χαρακτήρων που παρουσιάζουν περιορισμούς για χρήση σε πολυγλωσσικά υπολογιστικά συστήματα. Με τον κώδικα ASCII και τους άλλους κώδικες των 8 bit μπορούμε να αναπαραστήσουμε συνολικά μόνο 256 διαφορετικούς χαρακτήρες. Οι πρώτες 128 θέσεις δεσμεύονται για το Λατινικό αλφάβητο και μερικά σύμβολα, ενώ οι υπόλοιπες 128 για το τοπικό μη-Λατινογενές αλφάβητο.



Στο παρελθόν, διάφοροι οργανισμοί, συγκέντρωσαν διαφορετικά σύνολα χαρακτήρων και δημιούργησαν κωδικοποιήσεις γι' αυτά (π.χ. ένα σύνολο μπορεί να κάλυπτε μόνο τις λατινογενείς δυτικοευρωπαϊκές γλώσσες ή μια ιδιαίτερη γλώσσα της Άπω Ανατολής, όπως τα Ιαπωνικά). Αυτό είχε ως αποτέλεσμα μια εφαρμογή να μην μπορεί να υποστηρίξει όλες τις κωδικοποιήσεις, αλλά ούτε μια συγκεκριμένη κωδικοποίηση επαρκούσε για να καλύψει όλα τα γράμματα, σημεία στίξης και τεχνικά σύμβολα μιας συγκεκριμένης γλώσσας. Επιπλέον, ήταν συνήθως αδύνατος συνδυασμός διαφορετικών κωδικοποιήσεων στην ίδια ιστοσελίδα ή σε μια βάση δεδομένων

- 
- Ο κώδικας Unicode παρέχει ένα μεγάλο, ενιαίο σύνολο χαρακτήρων που έχει ως στόχο να συμπεριλάβει όλους τους χαρακτήρες που απαιτούνται για κάθε σύστημα γραφής στον κόσμο, συμπεριλαμβανομένων των αρχαίων συστημάτων γραφής (σφηνοειδής γραφή, γοτθική και αιγυπτιακή ιερογλυφική γραφή). Επίσης περιλαμβάνει και άλλα σύμβολα που χρησιμοποιούνται στα μαθηματικά, τις Φυσικές Επιστήμες και τη μουσική.
 - Πλέον, ο ρόλος του είναι σημαντικός στην αρχιτεκτονική του παγκόσμιου ιστού και τα λειτουργικά συστήματα, και υποστηρίζεται από όλα τα προγράμματα πλοήγησης και τις εφαρμογές.
 - Προτείνει έναν **μοναδικό αριθμό (code point)** για κάθε χαρακτήρα, ανεξάρτητα από το λειτουργικό σύστημα, το λογισμικό και τη φυσική γλώσσα. Οι πρώτες 65.536 ($=2^{16}$) θέσεις κωδικών σημείων (code points) στο σύνολο χαρακτήρων Unicode αποτελούν το Βασικό Πολυγλωσσικό Επίπεδο (Basic Multilingual Plane, BMP) και περιλαμβάνει τους χαρακτήρες που χρησιμοποιούνται περισσότερο. Η τιμή του code point παριστάνεται με το πρόθεμα U+ ακολουθούμενο από τη δεκαεξαδική μορφή της θέσης του. Για παράδειγμα στο code point U+0041 αντιστοιχεί το "Latin Capital letter A".

3.1.1 Κώδικας χαρακτήρων ASCII

- Προτείνει έναν **μοναδικό αριθμό (code point)** για κάθε χαρακτήρα, ανεξάρτητα από το λειτουργικό σύστημα, το λογισμικό και τη φυσική γλώσσα. Οι πρώτες 65.536 ($=2^{16}$) θέσεις κωδικών σημείων (code points) στο σύνολο χαρακτήρων Unicode αποτελούν το **Βασικό Πολυγλωσσικό Επίπεδο** (Basic Multilingual Plane, BMP) και περιλαμβάνει τους χαρακτήρες που χρησιμοποιούνται περισσότερο. Η τιμή του code point παριστάνεται με το πρόθεμα U+ ακολουθούμενο από τη δεκαεξαδική μορφή της θέσης του. Για παράδειγμα στο code point U+0041 αντιστοιχεί το "Latin Capital letter A".
- Παρέχει επίσης χώρο για περίπου ένα εκατομμύριο επιπλέον θέσεις κωδικών σημείων (code points) για **συμπληρωματικούς χαρακτήρες** (supplementary characters).



Ένα σύνολο χαρακτήρων, πολλαπλές κωδικοποιήσεις:

Πολλά πρότυπα κωδικοποίησης χαρακτήρων, όπως εκείνα της σειράς ISO 8859, χρησιμοποιούν ένα μόνο byte για ένα χαρακτήρα και η κωδικοποίηση είναι μια απλή αντιστοίχιση της θέσης του χαρακτήρα στο κωδικοποιημένο σύνολο χαρακτήρων με τον αντίστοιχο δυαδικό αριθμό. Για παράδειγμα, το λατινικό γράμμα A αντιστοιχίζεται στον αριθμό 65 που στο δυαδικό σύστημα αρίθμησης είναι 01000001 (1 byte). Για το πρότυπο ISO 8859-1 αυτό δεν αλλάζει ποτέ.



○ 𐀀 ○ 𐀁 𐀂 𐀃 !

2D30 2D63 2D53 2D4D 21

Character encodings

UTF-16

2D 30 2D 63 2D 53 2D 4D 00 21

UTF-8

E2 B4 B0 E2 B5 A3 E2 B5 93 E2 B5 8D 21

UTF-32

00 00 2D 30 00 00 2D 63 00 00 2D 53 00 00 2D 4D 00 00 00 21



□ Η κωδικοποίηση χαρακτήρων UTF-8 (8-bit Unicode Transformation Format) αποτελεί μια κωδικοποίηση μεταβλητού μήκους. Χρησιμοποιεί ομάδες από byte (από ένα μέχρι τέσσερα byte) για να αναπαραστήσει τα κωδικά σημεία (code points) του Unicode. Για παράδειγμα χρησιμοποιεί μόνο ένα byte για την κωδικοποίηση των 128 ASCII χαρακτήρων στο διάστημα του Unicode U+0000 μέχρι U+007F, δύο bytes για χαρακτήρες σε πολλές ομάδες αλφάβητων, και τρία bytes για τους υπόλοιπους χαρακτήρες του Βασικού Πολυγλωσσικού Επίπεδου (BMP). Για τους Συμπληρωματικούς Χαρακτήρες (supplementary characters) - που σπάνια χρησιμοποιούνται- χρησιμοποιεί 4 bytes.

□ Η κωδικοποίηση UTF-16 χρησιμοποιεί 2 bytes για κάθε χαρακτήρα στο Βασικό Πολυγλωσσικό Επίπεδο (BMP), και 4 byte για τους Συμπληρωματικούς Χαρακτήρες.

□ Η κωδικοποίηση UTF-32 χρησιμοποιεί 4 byte για όλους τους χαρακτήρες

Το UTF-8 (8-bit Unicode Transformation Format) είναι ένα μη-απωλεστικό σχήμα κωδικοποίησης χαρακτήρων μεταβλητού μήκους για το πρότυπο Unicode. Χρησιμοποιεί ομάδες από byte για να αναπαραστήσει τα κωδικά σημεία του Unicode.

Συγκεκριμένα χρησιμοποιεί ένα μέχρι τέσσερα byte ανά χαρακτήρα ανάλογα με το σύμβολο και το κωδικό του σημείο. Για παράδειγμα χρειάζεται μόνο ένα byte του UTF-8 για την κωδικοποίηση των 128 ASCII χαρακτήρων στο διάστημα του Unicode U+0000 μέχρι U+007F.

Εύρος κωδικών δεκαεξαδικό	UTF-8 δυναδικό
000000-00007F	0
000080-0007FF	110xxxxx 10xxxxxx
000800-00FFFF	1110xxxx 10xxxxxx 10xxxxxx
010000-10FFFF	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx

παράδειγμα

Για παράδειγμα ο χαρακτήρας άλεφ (**א**), που είναι ο χαρακτήρας Unicode με κωδικό σημείο U+05D0, μέσω του σχήματος UTF-8 κωδικοποιείται ως εξής:

Βρίσκεται στο διάστημα U+0080 - U+07FF. Ο πίνακας δείχνει ότι θα κωδικοποιηθεί χρησιμοποιώντας δύο byte **110xxxxx 10xxxxxx**.

Ο δεκαεξαδικός 0x05D0 είναι ισοδύναμος στον δυαδικό 101-1101-0000.

Τα έντεκα bit τοποθετούνται με την σειρά που έχουν στις θέσεις που σημειώνονται με "x": **11010111 10010000**.

Το τελικό αποτέλεσμα είναι δύο byte, που για οικονομία τα εκφράζουμε ως δύο byte στο δεκαεξαδικό 0xD7 0x90. Αυτό είναι το γράμμα άλεφ στο σχήμα UTF-8.



v

v Greek Small Letter Nu

U+03BD

ς

ς Greek Small Letter Final Sigma

stigma (the modern greek name for this letterform)

U+03C2

IV

IV Roman Numeral Four

U+2163



☎ Black Telephone Emoji

rotary phone

U+260E



ν

ν Greek Small Letter Nu

U+03BD

11001110 10111101

ς

ς Greek Small Letter Final Sigma

stigma (the modern greek name for this letterform)

U+03C2

11001111 10000010

IV

IV Roman Numeral Four

U+2163

11100010 10000101 10100011

☎ Black Telephone Emoji

rotary phone

U+260E

11100010 10011000 10001110





Σας ευχαριστώ για την προσοχή σας