

ΚΩΔΙΚΟΠΟΙΗΣΗ UTF-8

1. Επισκέψου την ιστοσελίδα <https://unicode-table.com/en/sets/>, η οποία περιέχει τον πίνακα χαρακτήρων unicode.
2. Αφού μεταβείς στους ελληνικούς χαρακτήρες, επέλεξε έναν από αυτούς.
3. Σημείωσε τον αριθμό unicode, και κατόπιν μετέτρεψε τον στην μορφή με την οποία θα αποθηκευτεί στο εσωτερικό του υπολογιστή σε μορφή UTF-8..

Θυμήσου:

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 δυαδικό	Σημειώσεις
000000-00007F	0xxxxxxx	εύρος ισοδυναμίας με το πρότυπο ASCII, το byte αρχίζει με 0
000080-0007FF	110xxxxx 10xxxxxx	το πρώτο byte αρχίζει με 110 ή 1110, το επόμενο ή επόμενα byte αρχίζουν με 10
000800-00FFFF	1110xxxx 10xxxxxx 10xxxxxx	
010000-10FFFF	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx	

Για παράδειγμα ο χαρακτήρας **άλεφ (Α)**, 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": 110**10111** 100**10000**.
- 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.