Κεφάλαιο 3ο - Αναπαράσταση αλφαριθμητικών χαρακτήρων και αριθμών

3.1 Κωδικοποίηση χαρακτήρων

Οι λέξεις και οι φράσεις σε ένα κείμενο, δημιουργούνται από χαρακτήρες. Οι χαρακτήρες ομαδοποιούνται σε **σύνολα χαρακτήρων (character sets ή repertoires)**. Ένα σύνολο χαρακτήρων μπορεί να περιλαμβάνει αλφαβητικούς χαρακτήρες, αριθμούς και άλλα σύμβολα. Ένα σύνολο χαρακτήρων λέγεται **κωδικοποιημένο σύνολο χαρακτήρων** όταν κάθε χαρακτήρας αντιστοιχίζεται σε ένα συγκεκριμένο αριθμό που καλείται **code point**. Η τιμή του code point, αντιπροσωπεύει τη θέση του χαρακτήρα στο κωδικοποιημένο σύνολο χαρακτήρων.
*Για παράδειγμα, στο κωδικοποιημένο σύνολο χαρακτήρων Unicode, το code point για το λατινικό χαρακτήρα **A** είναι το 45στο δεκαδικό σύστημα αρίθμησης.*
Τα κωδικοποιημένα σύνολα χαρακτήρων, μερικές φορές αναφέρονται ως **κωδικοσελίδες (code pages)**.
Η **κωδικοποίηση χαρακτήρων (character encoding)** είναι το κλειδί που μετατρέπει τα code points σε δυαδική μορφή για να αποθηκευτούν σε έναν υπολογιστή ή να μεταδοθούν.
Επομένως τα μέσα και οι συσκευές εισόδου, εξόδου, αποθήκευσης και επικοινωνίας μεταφέρουν, αποθηκεύουν και χειρίζονται δεδομένα κειμένου υπό τη μορφή ενός κώδικα χαρακτήρων που απεικονίζει κάθε χαρακτήρα σε έναν μοναδικό αριθμό (code point). Η κωδικοποίηση μετατρέπει αυτούς τους αριθμούς σε δυαδική μορφή, ώστε κάθε χαρακτήρας να αντιστοιχίζεται σε μία μοναδική διαφορετική ακολουθία δυαδικών ψηφίων (σειρές από bytes).
Οι πιο διαδεδομένοι κώδικες για την αναπαράσταση χαρακτήρων είναι οι **κώδικες ASCII και UNICODE**

Κώδικας χαρακτήρων ASCII

Ο κώδικας **ASCII (American Standard Code for Information Interchange, Αμερικανικός Πρότυπος Κώδικας για Ανταλλαγή Πληροφοριών)** είναι ένα κωδικοποιημένο σύνολο χαρακτήρων του λατινικού αλφάβητου. Υποστηρίζει την αναπαράσταση των 256 χαρακτήρων (2^8=256) Οι πρώτες 128 θέσεις δεσμεύονται για το Λατινικό αλφάβητο και μερικά σύμβολα, ενώοι υπόλοιπες 128 για το τοπικό μη-Λατινογενές αλφάβητο

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 υπερσύνολο όλων των άλλων κωδικοποιήσεων συνόλων χαρακτήρων που παρουσιάζουν περιορισμούς για χρήση σε πολυγλωσσικά υπολογιστικά συστήματα.

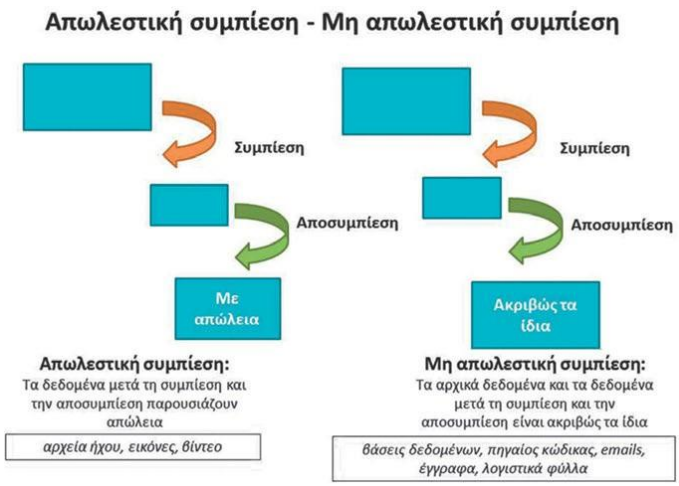
Ο κώδικας Unicode παρέχει ένα μεγάλο, ενιαίο σύνολο χαρακτήρων που έχει ως στόχο να συμπεριλάβει όλους τους χαρακτήρες που απαιτούνται για κάθε σύστημα γραφής στον κόσμο, συμπεριλαμβανομένων των αρχαίων συστημάτων γραφής *(σφηνοειδής γραφή, γοτθική και αιγυπτιακή ιερογλυφική γραφή)*. Επίσης περιλαμβάνει 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".*
Παρέχει επίσης χώρο για περίπου ένα εκατομμύριο επιπλέον θέσεις κωδικών σημείων (code points) **για συμπληρωματικούς χαρακτήρες (supplementary characters)**.
Ένα σύνολο χαρακτήρων, πολλαπλές κωδικοποιήσεις:

Η κωδικοποίηση χαρακτήρων **UTF-8 (8-bit Unicode Transformation Format)** αποτελεί μια κωδικοποίηση μεταβλητού μήκους. Χρησιμοποιεί ομάδες από byte (από 1 μέχρι 4 byte) για να αναπαραστήσει τα κωδικά σημεία (code points) του Unicode.
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

Μέθοδοι συμπίεσης δεδομένων
Η **συμπίεση δεδομένων (data compression)**, είναι η διαδικασία περιορισμού του μεγέθους ενός αρχείου, ώστε να χρειάζεται λιγότερο χώρο για την αποθήκευση ή τη μετάδοσή του.
Για παράδειγμα, αν μειώσεις κατά το ήμισυ το μέγεθος ενός αρχείου, μπορείς να αποθηκεύσεις περισσότερα αρχεία στο ίδιο αποθηκευτικό μέσο, ή να κατεβάσεις (download) τα αρχεία πιο γρήγορα (και στο ήμισυ του κόστους εάν πληρώνεις για το κατέβασμα).
Παρόλο που η χωρητικότητα των δίσκων ή η ταχύτητα σύνδεσης στο Διαδίκτυο συνεχώς αυξάνει, με μικρότερα συμπιεσμένα αρχεία υπάρχει πάντα κέρδος (χώρος, χρόνος, κόστος). Η διαδικασία της συμπίεσης εφαρμόζεται συστηματικά στα υπολογιστικά συστήματα που χρησιμοποιούν και επεξεργάζονται μεγάλο όγκο ψηφιακών δεδομένων *(π.χ. Google, Facebook)*, όπου μείωση του χώρου αποθήκευσης σημαίνει μείωση των υπολογιστών που απαιτούνται και κατά συνέπεια μεγάλη εξοικονόμηση ενέργειας και προστασία του περιβάλλοντος.

Τα συμπιεσμένα δεδομένα για να επεξεργαστούν, πρώτα **αποσυμπιέζονται (decompression)**.

Η μέθοδοι συμπίεσης χωρίζεται σε **απωλεστικές**, όταν έχουμε κάποια μικρή απώλεια της αρχικής πληροφορίας και σε **μη απωλεστικές,** όταν το τελικό αρχείο μετά την αποσυμπίεση είναι αντίγραφο του αρχικού.



Συμπίεση κειμένου με τον αλγόριθμο κωδικοποίησης Ziv-Lempel
Η βασική ιδέα της **κωδικοποίησης Ziv-Lempel** είναι ότι σε πολλά αρχεία (κυρίως κειμένου), επαναλαμβάνονται συχνά ακολουθίες χαρακτήρων (για παράδειγμα το άρθρο «τον»).