Šta je kod i zašto je bila potrebna standardizacija

Kodovi su uvedeni da bi se prikaz slova i znakova pomoću binarnog sistema u racunarima standardizovao i da bi se slova zapisivala isto u svakom kompjuteru (bilo je potrebno zbog prenosa kodova sa jednog računara na drugi).

Kodovi određuju odnos između skupova bitova i znakova pisanog jezika, i omogućavaju digitalnim uređajima da međusobno komuniciraju i obrađuju i čuvaju podatke koje sadrže znakove.

Neki fontovi su sadržali prepravljene tabele, tako da su se umesto engleskih karaktera koji se ređe koriste ubacivala naša slova (š, đ, č, ć i ž). Pri prenosu dokumenta na drugi računar, naša slova su se gubila, osim ako ne bismo odneli i instalirali font koji smo koristili. Problem je bio I što oznake na tasterima tastature nisu postojale ni po jednom od naših bivših "standarda". Haos koji je nastao iz upotrebe ovog kvazi-standarda su hiljade različitih oblika slova, a išlo se dotle da su pojedini korisnici imali svoj lični font. Jedina opravdana upotreba je bila u pripremi grafičkog materijala gde se priprema za štampu obavljala na lokalnom računaru, a na njemu nije postojao nijedan sličan font po Unicode standardu. U vreme nepostojanja razvijenog standarda u praksi zaživelo je čak desetak različitih kodnih rasporeda, pa su korisnici gubili dragoceno vreme prilagođavajući setove karaktera pri razmeni dokumenata.

Na početku svi su se trudili da se uguraju u preostalu polovinu; svaki set simbola je bio standardizovan pod određenim nazivom. Jednog dana je postojalo više standarda nego simbola u bajtu i svi su se bitno razlikovali u delovima van prvih 128 simbola. Tako je došlo do potpunog haosa, računari su prikazivali nečitljive znakove i simbole…

Osim toga, postojao je i ogromni broj drugih kodova; došlo je i do različitih nikako ne povezanih kodiranja istih stvari, na primer MS DOS za ćirilicu koristi kodove 855 i 866, Windows koristi 1251, a Mac Os još nešto. Osim toga postoje i KOI 8 i KOI 7, pa i ISO 8859-5, i svako slovo ima više različitih izgleda.

Osnovni problem sa kodnim stranama je to sto se medjusobno iskljucuju, tj. cijeli dokument mora da bude napisan istim pismom.

ASCII

ASCII (American Standard Code for Information Interchange)(Američki standardni kod za razmenu podataka) je skup znakova i kodna stranica utemeljena na latinskom pismu kakvo koristi engleski jezik. Kod nas je popularan kao ošišana latinica.

Početak razvijanja ASCII koda bio je 1960. godine na sastanku udruženja ASA (American Standard Assosiation, kasnije nazvanog ANSI, te se ovaj kod naziva i ANSI_X3.4-1968). ASCII kod je zasnovan na tadašnjem telegrafskom kodu od strane Bell laboratorija. Kod se menjao kroz vreme, objavljen je 1963. godine, da bi se 1986. godine formirala njegova verzija koja se i danas koristi.

ASCII je na početku bio 7-bitni kod, popunjen je sa 100 karaktera, a ostalih 28 ostavljeni su kao mogućnost za dodavanje novih karaktera. Sadrži dve grupe karaktera: kontorlni (control characters) i vidljivi (printable characters). Kontrolni karakteri služili su za zapisivanje naredbi kao što su novi red, kraj teksta itd., dok su vidljivi služili za samo ispisivanje teksta. Kako su razvijani računari, prostor za čuvanje podataka postao je jedan bajt (8 bitova) i osmi bit je bio obično korišćen kao bit parnosti za proveru grešaka u prenosu podataka ili je imao ulogu karakterističnu za dati uređaj. Takođe, vremenom su nastale i razne verzije ovog koda kao što su YUSCII (Yugoslav Standard Code for Information Interchange), ISCII (Indijska verzija), VISCII (Vijetnam) itd.

1. Osnovni ASCII kod je sedmobitni kod, što znači da se njime može prikazati 128 znakova (2^7) tj. koristi sedam binarnih cifara (0-127 u dekadnom). Prva 32 (0-31) koda u tabeli ASCII-kodova su rezervisana za kontrolne (upravljačke) znakove (upravljaju štampačima,skenerima...). Kodovi od 33 do 126 (ukupno 95 karaktera) su oni koji se mogu štampati (od 20_{16} do $7E_{16}$) predstavljaju slova,cifre,znakove interpunkcije... (20_{16} je “SPACE” karakter i ulazi u karaktere koji se mogu štampati).

2. Prošireni ASCII kod koristi 8 bitne zamjene te može prikazati 256 (2^8) različitih znakova. Prvih 128 znakova jednako je standardnom ASCII kodu. Drugih 128 služi za prikaz posebnih znakova drugih jezika (dodati su još neki karakteri kao što su ćirilična slova, nemačka slova sa umlautima itd.) Zbog razlike u jezicima u različitim zemljama su donesene lokalne norme.

Za razliku od ranijih kodova, ASCII je organizovan tako da bude što lakše sortirati liste alfabetički.
Sva mala i velika slova u standardnom ASCII kodu se razlikuju samo po vrednosti petog bita, kod velikih slova on je 0 a kod malih 1. Na primer: G je 1000111, g je 1100111. Mala slova se mogu pretvoriti u velika (i obrnuto) promenom petog bita.
Bit 5 i 6 oznacavaju kojoj grupi karaktera pripada:

Bit 6	Bit 5	Grupa
0	0	Kontrolni karakteri
0	1	Cifre i interpunkcija
1	0	Velika slova i specijalni karakteri
1	1	Mala slova i specijalni karateri

U početku druge varijante ASCII-a su počele da se prave kako bi se osim engleskog izrazili drugi jezici.Prvi takav pokušaj se dogodio 1972. godine kada je napravljen ISO 464.Kasnije razvojom tehnologije razvili su se osmobitmni standardi kao što je ISO/IEC 8859 koji su predstavljali proširenje ASCII-a.Unikod i ISO/IEC 8859 imaju daleko širi izobr znakova i oni ubrzo zamenjuju ISO/IEC 8859 i ASCII.Može se reći da su oni proširenje ASCII-a. Većina modernih načina kodiranja karaktera su bazirani na ASCII, iako sadrže mnogo dodatnih karakera. Ruski standard KOI7 je modifikacija ASCII standarda takva da su sva mala slova zamenjena velikim ćiriličnim, pa nisu postojala mala slova

Kontrolni karakter, ili neštampajući karakter jeste karakter koji ne predstavlja simbol za pisanje. U ASCII tabeli to su karakteri do koda 32, a i kod 127 (DEL karakter) je takodje kontrolni karakter. U novijim načinima kodiranja, kao što su ISO-8 i UNICODE, postoji još kontrolnih karaktera.

Ispod su navedeni neki od bitnijih kontrolnih karaktera iz ASCII tabele i njihova značenja:

- 0 (null, NUL ,\0, ^@), prvobitno je trebalo da bude karakter koji bi se potpuno ignorisao, ali ga u današnje vreme mnogi programski jezici koriste za označavanje kraja stringa;
- 7 (bell, BEL, \a, ^G), često se koristi da natera uredjaj da pošalje upozorenje (najčešće zvučno);
- 8 (backspace, BS, \b, ^B), koristi se da obriše poslednji ispisani karakter;
- 9 (horizontal tab, VT, \v, ^K), pomera sve karaktere u redu za nekoliko mesta udesno;
- 10 (line feed, LF, \n, ^J), označava kraj reda na većini UNIX sistema;
- 12 (form feed, FF, \f, ^L), daje informaciju štampaču da izbacuje papir do početka sledeće stranice ili terminalu da očisti ekran;
- 13 (carriage return, CR, \r, ^M), vrlo je sličan line feed-u s tim što resetuje poziciju na početak reda;
- 26 (Control-Z, SUB, EOF, ^Z,) koristi se kao zamena za nepoznati karakter;
- 27 (escape, ESC, \e, ^[), ima razne primene, najčešće se koristi da prekine neki proces;
- 127 (delete, DEL, ^?), kao i 0, planirano je da se koristi kao karakter koji se ignoriše, ali se sada u nekim sistemima koristi da izbriše karakter.

YUSCII

Onog trenutka kada su računari zaživeli po celom svetu, a ne samo na engleskom govornom području, javili su se problemi sa nacionalnim znakovima. ASCII je već bio popunjen, pa je jedno od prvih rešenja bilo da se pojedini simboli "žrtvuju". Tako su kodne reči rezervisane u ASCII-ju za '@', '[', '\', ']', '^', '`', '{', '|', '}', '~' kod nas pridružene simbolima 'Ž', 'Š', 'Đ', 'Ć', 'Č', 'ž', 'š', 'đ', 'ć', 'č' i tako je nastao YUASCII ili YUSCII.

YUSCII je 7-bitni kod zasnovan na ISO 646. Korišćen je u Jugoslaviji pre nego što su ušli u upotrebu ISO 8859-2, Windows -1250/1251 i Unicode. Svoje korene vuče još iz DOS-a, iako je DOS imao predviđen kodni raspored za naše podneblje po imenu CP 852.

Kodna strana 852 (poznata kao CP 852, IBM 00852, OEM 852 (Latin II), MS-DOS Latin 2) je kodna strana korišćena pod DOS-om za pisanje na centralno evropskim jezicima (kao što su srpski, hrvatski, češki, mađarski, poljski, rumunski i slovački).

Kodna strana 852 (DOS Latin 2) je vrlo različita od ISO/IEC 8859-2 (ISO Latin-2), iako se obe vode kao "Latin-2", doduše u različitim jezičkim područjima.

ISO-8

ISO-8 (International Organization for Standardization-8) je osmobitni kod koji lici na EASCII (Extended ASCII), dakle sadrži 256 karaktera, od kojih se prvih 127 poklapa sa ASCII kodom.

ISO-8859-2

ISO-8859-2 je 8-bitni kod. Prvih 128 karaktera su identični sa ASCII karakterima i obično se nazivaju donja kodna strana. Naša slova se nalaze među preostalih 128 karaktera u tzv. gornjoj kodnoj strani.

EBCIDIC

EBCDIC (*Extended Binary Coded Decimal Interchange Code)* je standardni 8-bitni kod koji je stvorio IBM 1963. godine za svoju proizvodnju. Koristio se na *IBM* mainframe z/OS, s/390, AS/400 i i5/OS. Razvio se od 6-bitnog BCD sistema za kodiranje bušenih kartica. U ovo vreme EBCDIC bio je opšte nepopularan među programerima jer je postojalo bar šest verzija ovog koda koje su nekompatibilne međusobno, a ni jedna verzija ne sadrži sve karaktere koji su potrebni za moderno programiranje. Napravljen je kao proširenje BCDIC sistema za kodiranje (Binary Coded Decimal Interchange Code). Za ovaj kod ima dosta šala zato što je napravljen po zahtevu vlade Amerike. Jedna od njih ide ovako:*Professor: "So the American government went to IBM to come up with an encryption standard, and they came up with—"Student: "EBCDIC!"*

IBM PC

IBM PC je takođe razvio osmobitni kod, kod koga se prvih 127 karaktera poklapa sa ASCII kodom, a ostatak je originalno definisan.

UNICOD

Ideja UNICODE (UNIversal enCODE) je bila jednostavna. Svaki simbol je dobijao svoj kod jednom i zauvek. Nastao je 1990. godine da bi se omogućilo ispisivanje teksta na više različitih jezika tj da se dogovori jedan standard za sve karaktere koji postoje. Sadrži više od 128.000 različitih slova. Karakteri se u kompjuteru mogu prikazati kao 8 bitova, 16 bitova ili 32 bita. Postoje dvije organizacije koje definišu dva standarda za Unicode. Jedan format je razvijen od strane tzv. The Unicode Consortium pod nazivom The Unicode Standard. Drugi standard je razvila Međunarodna organizacija za standardizaciju - International Organization for Standardization. Ta dva standarda su skoro identična i razlikuju se po predstavljanju kineskih, japanskih I korejskih znakova.

Postoji više verzija Unikoda. Osnovna verzija je dvobajtni format zapisa do 2^{16} = 65536 znakova. Njen naziv je USC -2 zato što koristi dva okteta, odnosno dva bajta. Sa 16 bitova (65536 karaktera) je rešen problem zapisa svih postojecih pisama (uključujući čak i neka izmišljena, kao na primer klingonsko pismo). Postoji i noviji UNICODE standard pod nazivom UCS-4 koji koristi 4 bajta za zapis 2^{31} = 2147483648 znakova podeljenih u tzv. ravni. Prva dva bajta definišu ravan, tako da ima 2^{15}= 32768 ravni. Druga dva bajta definišu znak unutar ravni, tako da ima 2^{16} = 65536 znakova po ravni. Taj noviji format je više napravljen kao plan za budućnost nego kao realna opcija.

Prvih 256 Unicode-kodova se poklapaju sa karakterima u ISO-8859-1 i nazivaju se kao Latin -1 podskup Unicode-a. Naša slova pripadaju narednom delu od 128 karaktera, Latin Extended – A podskup Unicode-a. Jedan od problema je što nije uzeto u obzir da se karakteri nekih ćiriličnih pisama razlikuju, što se najčešće ispoljava u Italic-u.

Svjetski trendovi razvoja baza podataka idu ka uvođenju Unicode-a, kao standardni način zapisa podataka. Većina baza podataka već duže vreme podržava Unicode. Јуникод конзорцијум је међународно тело које се бави стандардима за комуникацију на дигиталним платформама, а надлежно је и за стандардизацију сличица којима нове генерације воле да комуницирају и изражавају своје емоције. Ускоро ће бити објављен Јуникод 9.0 у којем ће бити укључено 70-ак нових емотикона. Najnovija verzija je Unicode 9.0, sadrži malo više od 128000 karaktera.

UTF-8

UTF-8 (Unicode Transformation Format) je standard sposoban da zapiše sve karaktere kao Unicode, ali da zauzima manje prostora za češće karaktere. To je način kodiranja Junikoda, pri čemu se svaki UNICODE simbol kodira pomoću promenljivog broja (1-4) okteta (svaki oktet ima 8 bitova), pri čemu broj okteta zavisi od celobrojne vrednosti koja je dodeljena UNICODE karakteru. On je kompatibilan sa ASCII kodom. Razvili su ga Ken Thompson (razvio UNIX operativni sistem) i Rob Pike. Danas je najkorišćeniji od svih kodova.

Prvo je razvijena unicode transformaciona šema sa osnovnom jedinicom od 8 bita. Pomoću nje se karakter zapisuje u jednom, dva ili tri bajta, u zavisnosti od toga o kom je karakteru reč. U UTF-8 se karakter zapisuje u obliku jednog bajta ako u svom zapisu sadrži samo najnižih 7 bita, odnosno, ako je reč o ASCII karakteru Ukoliko karakter u svom Unicode zapisu sadrži samo najnižih 11 bita, u UTF-8 se zapisuje u obliku dva bajta. I na kraju, ako karakter sadrži svih 16 bita, zapisuje se u obliku tri bajta.

UTF-8 je zamišljen kao format koji najviše odgovara latiničnom tekstu. To je veoma pogodno za korišćenje u izvornom kodu programa ili u raznoraznim markup jezicima (HTML, XML, \LaTeX, ...) jer su standardne komande tih programskih/markup jezika uvek ASCII, a tekst koji se koristi može da bude i ASCII i UTF-8. Na taj način se ne ometa rad programskog kompajlera ili parsera markup jezika, a omogućava se korišćenje višejezičke podrške.

UTF-8 nije optimalan način zapisa za kineski i japanski tekst jer umesto da se koriste dva bajta po karakteru, za takav tekst bi bilo korišćeno čak tri bajta po karakteru, ali to i nije toliko važno za nas. Za ćirilični tekst je, sa druge strane, sve jedno da li se koristi čisti UNICODE ili UTF-8, pošto se svaki ćirilični karakter zapisuje u obliku dva bajta i u jednom i u drugom formatu. Za nas je ipak optimalniji UTF-8 jer postoji mogućnost pisanja i ćirilicom i latinicom, pa ako u ćirilici već ne može da se izbegne upotreba dva bajta, u latinici se skoro svi karakteri zapisuju samo sa jednim bajtom (osim šđčćž).