Olivera Mihailović
Beograd, Narodna biblioteka Srbije

UDK 004.223

UPOTREBA ODGOVARAJUĆEG KARAKTER SETA

Zašto postoji toliko različitih karakter setova za svaki od svetskih jezika, kako ih računari koriste i razumeju?

Osnovni jezik računara (kaže se i maternji) je engleski. Zaista, najveći zamajac celokupne softverske industrije nalazi se i dalje u Sjedinjenim Američkim Državama. Kako se računari koriste širom sveta potrebno je prilagođavanje potrebama korisnika sa različitih jezičkih područja.

Istorijski, sve je počelo kada se ukazala potreba da se u memoriji računara predstave slova. Morala se usvojiti neka konvencija prema kome bi se slova označila brojevima. Iako je jedan bajt dovoljan za prikazivanje 256 karaktera u početku je to izgledalo previše.

Jedan od karakter setova je **ASCII** (American Standard Code for Information Interchang). Ovaj set je 7-bitni (2^7 = 128 karaktera) i u sebe uključuje interpunkcijske oznake, cifre i slova engleskog alfabeta. ASCII set dovoljan je za prikaz informacija na engleskom jeziku, međutim nedovoljan je za prikaz većine drugih jezika jer mnogi jezici sadrže takozvane specijalne karaktere, npr. apostrofe, dvotačke, akcente… Dakle, bilo je potrebno proširenje karakter seta kako bi opslužio sve jezike, u cilju obezbeđenja razmene informacija između kompjuterskih sistema. Srećom prostora za proširenje je bilo dovoljno, jer je 128 mesta bilo prazno. Tako je nastao standardni MS DOS set karaktera ili CP 437 (kodna strana 437). Sa pojavom MS DOS-a 5.0 *Microsoft* se ozbiljnije bavi sistemskom podrškom za nacionalne setove karaktera. Definisane su 8-bitne nove kodne stranice. Našoj latinici odgovara kodna strana 852, a ćirilici CP 855. Međutim, nije vođeno računa o činjenici da se na nekim prostorima koriste podjednako oba pisma .

Sa lansiranjem *Windows*-a, 3.1. 1992. godine, ključna novina je pojava *True Type* fontova u čiju je osnovu ugrađena višejezička podrška. Definisane su nove kodne strane. Izostavljeni su grafički simboli, ali su ubačena slova karakteristična za gotovo sve zapadnoevropske i skandinavske jezike. Međutim, pri osmišljavanju kodnih stranica *Microsoft* se nije do kraja pridržavao standarda koji je već ranije usvojila ISO. Pojavom *Windows*-a 95 *True Type* fontovi su prošireni da podržavaju kodne stranice 1250, 1251, 1252,1253, 1254 i 1257, tj. sadrže 652 karaktera – takozvani WGL4 (*Windows Glyph List* 4). Od pojave *Windows*-a NT može se koristi Unicode standard.

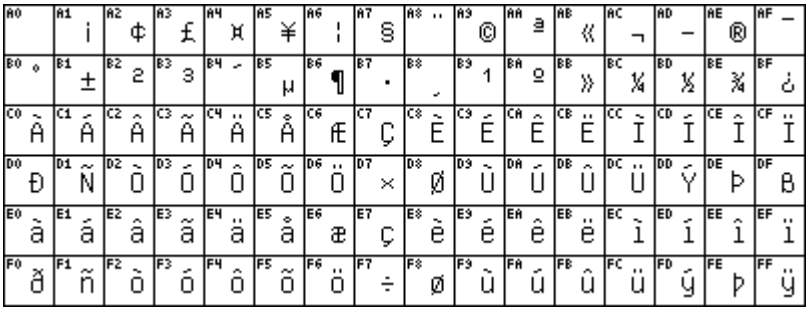
ISO-8859 karakter set dizajniran je sredinom 80-tih od European Computer Manufacture's Association (ECMA) i potvrđen od International Standards Organisation (ISO). Ovo je osmobitni karakter set – dakle sadrži 256 karaktera (2^8= 256). ISO 8859 karakter set je serija od 10 i više standardizovanih višejezičkih grafičkih karakter setova. Najvažniji su:

1. Latin1 (Zapadna Evropa)
2. Latin2 (Istočna Evropa)
3. Latin3 (Južna Evropa)
4. Latin4 (Severna Evropa)
5. Cyrillic
6. Arabic
7. Greek
8. Hebrew
9. Latin5
10. Latin6

„I'm really terrified to see how difficult it can be for a non-latin1 person to be able to print in his/her own mother tongue!" — Akim Demaille, 1998.

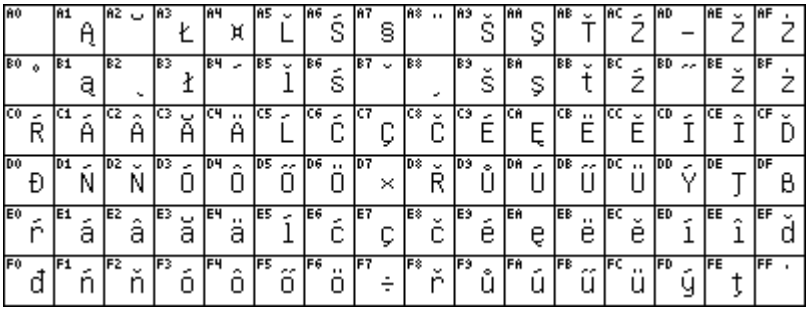
(Bio sam zaista zastrašen da vidim kako je teško „ne-latin1" osobi da može da štampa na svom maternjem jeziku!)

ISO-8859-1 (Latin1)



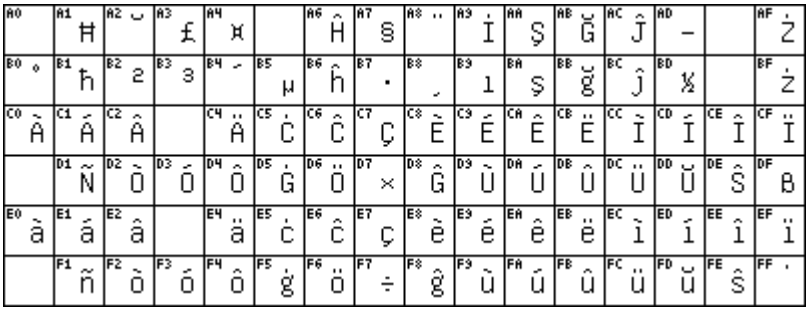
Karakter set „Latin1", poznat i kao zapadni karakter set (oznake ISO 8859-1) zadovoljava široke potrebe kao što je i sam Internet standard. Latin1 pokriva većinu zapadnoevropskih jezika : francuski, španski, portugalski, italijanski, albanski, nemački, norveški, finski, škotski, irski, engleski…

ISO-8859-2 (Latin2)



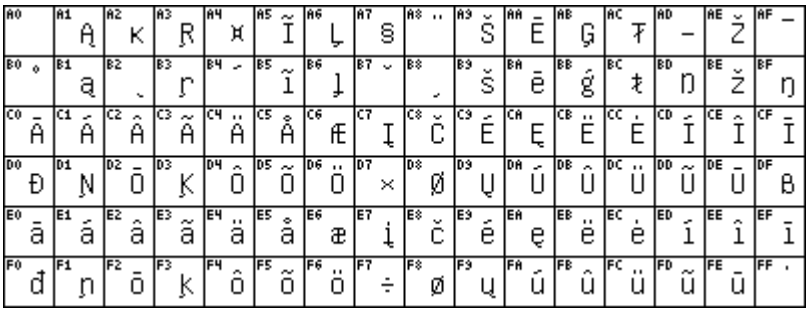
Latin2 pokriva centralnu i istočnu Evropu, odn. češki, mađarski, poljski, rumunski, hrvatski, slovački i slovenački jezik.

ISO-8859-3 (Latin3)



Latin3 je popularan kod autora esperanta, a koristio se za turski jezik pre uvođenja Latin5.

ISO-8859-4 (Latin4)

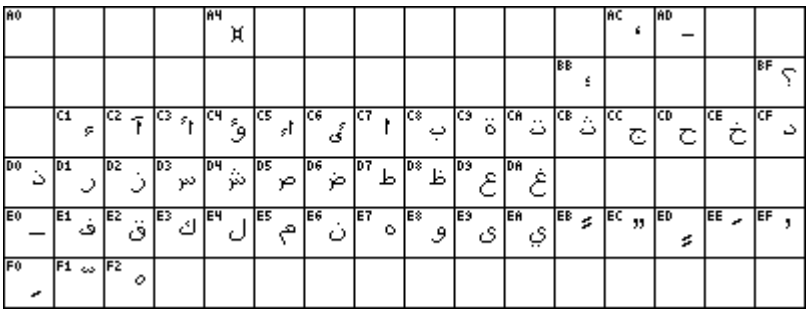


Latin4 predstavlja slova za estonski jezik.

ISO-8859-5 (Cyrillic)

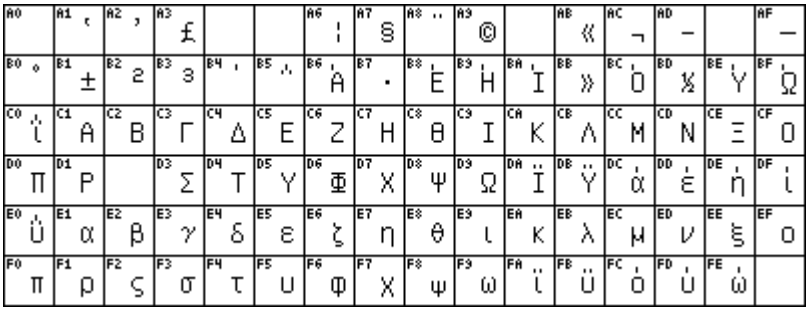
Ova ćirilična slova koriste se za bugarski, beloruski, makedonski, ruski i srpski jezik.

ISO-8859-6 (Arabic)



Ovo je arapski alfabet. Nažalost ne sadrži četiri eksta karaktera za persijski, ni osam eksta za pakistanski jezik.

ISO-8859-7 (Greek)

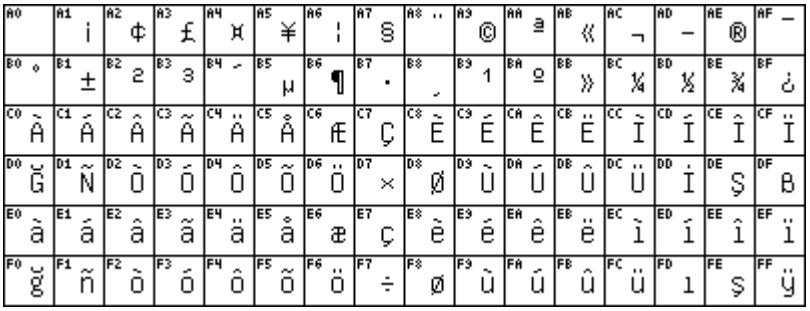


Koristi se za grčki jezik. ISO-8859-7 naziva se još ELOT-928 ili ECMA-118:1986.

ISO-8859-8 (Hebrew)

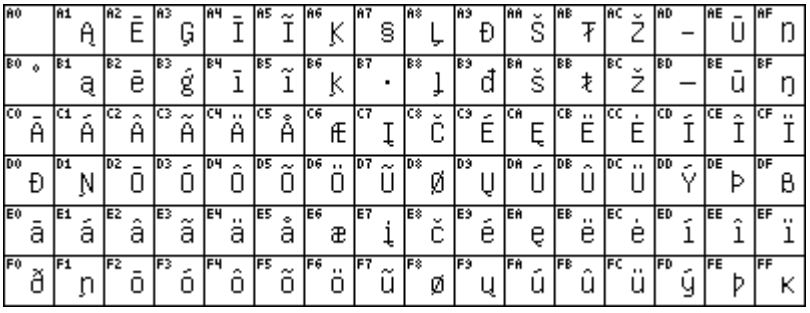
Koristi se za herbejski jezik i za jidiš.

ISO-8859-9 (Latin5)



Koristi se za turski jezik.

ISO-8859-10 (Latin6)



Pojavio se 1992. godine. Predstavlja reorganizaciju Latin4 karakter seta. Naziva se još i nordijski.

Po pravilu se ASCII karakteri pojavljuju kao podskup većine drugih karakter setova. Na primer, prvih 128 karaktera određenog seta predstavlja sve ASCII karaktere. Ovo je neophodno zato što HTML tagovi u web stranicama moraju ostati isti, bez obzira na to kojim jezikom se prikazuje stranica.

U slučaju kreiranja sajta koji će imati verziju i na srpskom i na engleskom jeziku, može se koristiti ISO 8859-1 karakter set, koji će zadovoljiti obe verzije.

Karakter setovi su potrebni prvenstveno radi konverzije kodne oznake u karaktere (character encoding). Browser-u mora biti naznačeno koja vrsta kodiranja se koristi, da bi svaku kodnu oznaku preveo u odgovarajući karakter. Upravo ovome i služi vrednost karakter seta unutar navedenog <META> taga u okviru HTML-a.

Windows ANSI je osmobitni karakter set dat kao set karaktera iznad i ispod 128 karatera. Donji deo tabele je identičan ASCII karater setu.

kodna strana iznad 128	1250 Istočna Evropa	1251 ćirilica	1252 zapadno evropski	1253 grčki ANSI	1254 turski	itd.
Ispod 128	ASCII	ASCII	ASCII	ASCII	ASCII	itd.

Unicode karakter set

Unicode standard određuje Unicode konzorcijum (Unicode Consortium - www.unicode.org). Isti standard obuhvaćen je i ISO specifikacijom, i to standardom ISO/IEC 10646.

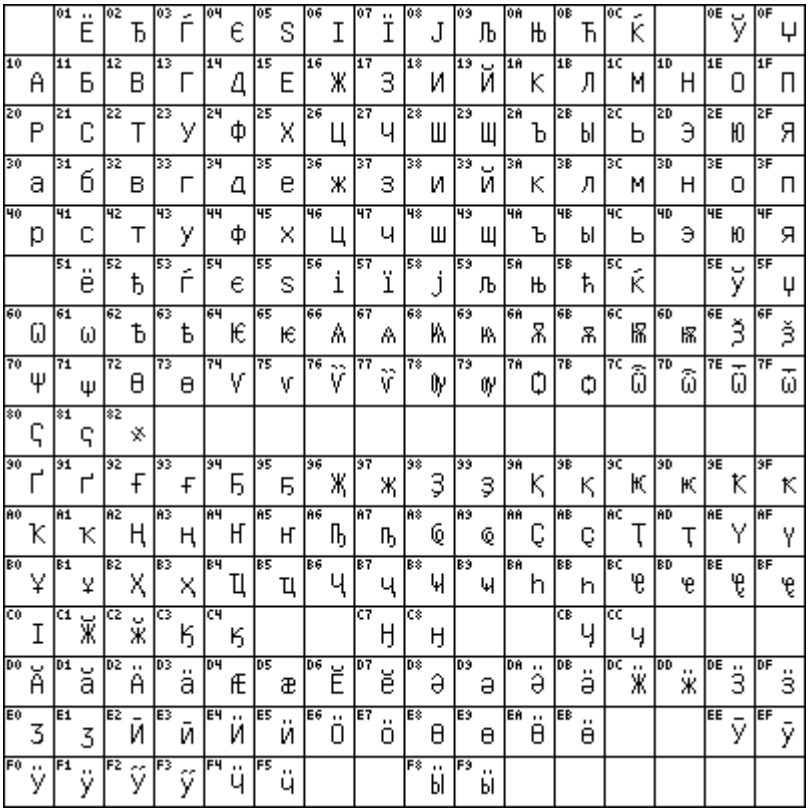
Unicode provides a unique number for every character,
no matter what the platform,
no matter what the program,
no matter what the language.

(Sa Unicode web sajta)

(Unicode dodeljuje jedinstven broj za svaki karakter
bez obzira na platformu
bez obzira na program
bez obzira na jezik)

Unicode predstavlja najvažniju ekstenziju ASCII karakter seta. Unicode karakter set podržava karaktere gotovo svih govornih jezika, kao i matematičke i druge simbole. Pored engleskog i ostalih zapadnoevropskih jezika, podržava i japanski i kineski, kao i našu ćirilicu. Podrška svih subjekata Unicode standardu predstavlja najvažniji korak u internacionalizaciji Interneta.

Sledeća tabela prikazuje U+0400 ćirilični blok Unicode tabele. Prati redosled ISO-8859-5.



Unicode fontovi:

Lucida Sans Unicode, od Bigelow & Holmes, isporučivan je sa nekim was verzijama. Sadrži 1775 simbola.

Times New Roman, *Arial*, *Courier i Impact* .Ovi fontovi sadrže WGL4 karakter set, koji podržavaju najveći deo latin-alfabeta, grčki i ćirilicu. Svaki sadrži 653 simbola, posebno Times New Roman sadrži malo iznenađenje sa 654-tim znakom.

Bitstream Cyberbit je realizovan kao TrueType font za Windows 95 i NT. Sadrži preko 8500 karaktera.Podržava većinu evropskih jezika uključujući grčki, ruski i turski. Takođe podržava japanski, korejski i kineski jezik.

TrueType GX font sadrži dosta karaktera, ali ne obavezno i Unicode tabelu.

Unicode je 16-bitni karakter set, što znači da svaki karakter sadrži dva bajta (DBCS - double byte character set). Na ovaj način moguće je mapirati 2^{16} = 65536 karaktera. Karakteri koji se koriste u XML dokumentima pripadaju Unicode karakter setu. Kako svaki Unicode karakter zauzima dvostruko više mesta od njegovog ekvivalenta Latin1 seta, i sam XML bi trebalo da je dva puta veći od normalnih tekstualnih fajlova. Ipak, u najvećem broju slučajeva nije potrebno svih 16 bita po karakteru, pa XML dokumente možemo formirati i sa 8-bitnim karakter setom. Pri tome, XML procesor mora da prepozna i *UTF-8* (osmobitni) set i *UTF-16* (šesnaestobitni) set.

Rezime

Kako je osnovni jezik računara engleski i kako se računari koriste širom sveta, ukazala se potreba prilagođavanja rada na računarima korisnicima sa različitih govornih područja. Morala se usvojiti neka konvencija prema kojoj bi se slova označila brojevima. Iako je jedan bajt dovoljan za prikazivanje 256 karaktera, u početku je to izgledalo previše.

Jedan od karakter setova je ASCII. Ovaj set je 7-bitni. ASCII set dovoljan je za prikaz informacija na engleskom jeziku, međutim nedovoljan je za prikaz većine drugih jezika, jer mnogi jezici sadrže takozvane specijalne karaktere. Dakle, potrebno je proširenje karakter seta kako bi opslužio sve jezike, obezbeđujući razmenu informacija između kompjuterskih sistema. Dalji razvoj dovodi nas do osmobitnih karakter setova ISO-8859, Windows ANSI i, najzad, do Unicode šesnaestobitnog karakter seta. Po pravilu se ASCII karakteri pojavljuju kao podskup većine drugih karakter setova.

Unicode predstavlja najvažniju ekstenziju ASCII karakter seta. Unicode karakter set podržava karaktere gotovo svih govornih jezika, kao i matematičke i druge simbole. Podrška svih subjekata Unicode standardu predstavlja najvažniji korak u internacionalizaciji Interneta.

Оливера Михаилович

Употребление подходящего набора знаков

Резюме

Английский язык является основным языком при пользованию компьютерами в целом мире и поэтому возникла потребность сделать работу на компьютере для пользователей, говорящих различные языки, более удобной. Должно было усвоить какую-то конвенцию, согласно которой могли бы буквы обозначить числами. Хотя один байт является достаточным для представления 256 знаков, в начале это выглядело слишком много.

Одним из наборов знаков является ASCII. Этот набор состоит из 7 битов. ASCII набор достаточен для показания информаций на английском языке, но недостаточен для показания большинства других языков, так как многие языки содержат, так называемые, специальные знаки. Значит, надо расширить набор знаков, чтобы он мог служить всем языкам, обеспечивая обмен информациями между компьютерными системамы. Дальнейшее развитие приводит нас к осмобитным наборам знаков ISO-8859, Windows ANSI и наконец до шестнадцатибитного набора знаков Unicode. Как правило ASCII знаки являются подгруппой большинства других наборов знаков.

Unicode представляет собой самое важное расширение ASCII набора знаков. Unicode набор знаков поддерживает знаки почти всех языков, как и математические и другие символы. Поддержка всех субъектов в Unicode стандарте представляет важнейший шаг в интернационализации Интернета.

Olivera Mihailović

Usage of Apropriate Character Set

Summary

Since the fact that English language has became the elementary language of/for the Computer Technology and as the usage of Personal Computers is worldwide now, it became apparent that there is a need for adaptation of the data processing on computers, especially for Users from different language zone. A Convention had to be adopted to replace letters with numbers. Although one byte is sufficient for the representation of 256 characters, at the beginning it seems too much.

One of the character set is ASCII (American Standard Code for Information Interchange). This set employs a 7-bit character code. ASCII set is adequate for displaying the information in English language, but not adequate enough to display the majority of the world's languages, because most of the other languages contain the so-called special characters. Therefore, it is necessary to expand the character of the set, so it can be suitable for all languages, providing the exchange of information between computer systems and interfaces. The further development lead us to ISO-8859, a character set of 8-bit, Windows ANSI and, at the end, to UNICODE, a character set of 16-bit. According to rules, the ASCII characters appear as subset of the major other character sets.

UNICODE represents the most important extension of the ASCII character set. UNICODE character set supports almost all spoken languages, as well as mathematics' and other symbols. The support of the UNICODE standards, for all subjects, represents the most important step in the internationalization of the Internet.