Filip Brčić
(Elektrotehnički fakultet, Beograd)

UNICODE

Dajemo pregled načina kodiranja višejezičkog teksta u elektronskoj formi koristeći standard Unicode, s posebnim osvrtom na varijantu UTF-8, koja je najzgodnija za kodiranje pretežno latiničnog teksta. Dajemo i kratko uputstvo za korišćenje te varijante u Microsoft Word-u, Netscape Composer-u i editoru teksta Kate. Takođe preporučujemo standardne Unicode fontove koji omogućuju laku prenosivost teksta s računara na računar ili za njegovo objavljivanje na Internetu.

1. Razvoj elektronskog zapisa teksta

Prvi računari su bili pravljeni pretežno za englesko govorno područje i imali su podršku samo za engleski alfabet, za brojeve, zagrade i još po neki kontrolni znak, što je činilo ukupno 128 mogućih slova (u 7 bita). To je bio tzv. ASCII ili US-ASCII standard. Kasnije je skup znakova proširen na 256 (8 bita), a ,,gornjih" 128 mesta je bilo korišćeno za dodatne znake. Iz navike je i ovaj prošireni ASCII nazivan ASCII, tako da tu često dolazi do zabune. Da bi postojala podrška za više jezika, smišljane su tzv. kodne strane (Code Pages) koje definišu ponašanje tog dodatnog skupa slova. Osnovna kodna strana na personalnim računarima (PC437) u tom gornjem skupu znakova definiše razne grafičke znake za crtanje tekstualnih prozora i slično. Kasnije je razvijeno još puno kodnih strana koje podržavaju određene jezike. Tako postoje Latin1 (ISO-8859-1) za latinična pisma Zapadne Evrope (Francuska, Nemačka, Španija, ...), Latin2 (ISO-8859-2) i Windows-1250 za latinična pisma Istočne Evrope (naša latinica i sl.), ISO-8859-5, KOI8-R i Windows-1251 za ćirilicu... Osnovni problem s kodnim stranama je to što se međusobno isključuju, tj. ceo dokument mora da bude napisan istim pismom. To uglavnom nije teško realizovati, ali ako bi bilo potrebno pomešati dva pisma, kao naprimer u nekom turističkom vodiču gde zajedno postoje tekst na srpskom, na engleskom i na francuskom, nailazi se na problem. Zbog toga se došlo do ideje da se napravi jedinstveni zapis za sve jezike – Unicode.

2 Pregled postojećih verzija Unicode-a

Postoji više verzija Unicode-a. Bazična verzija je dvobajtni format zapisa do 2^{16} = 65536 znakova. Njen naziv je UCS-2[1] zato što koristi dva okteta, odnosno dva bajta. Sa tih 65536 znakova rešen je problem zapisa skoro svih postojećih pisama (uključujući

[1] UCS = Universal Multiple-Octet Coded Character Set, što znači Univerzalan višeoktetno (tj. višebajtno) kodiran skup karaktera

čak i neka izmišljena, kao na primer Klingon-sko pismo). Ovaj tip Unicode-a se naziva Plain UCS-2 ili UTF-16.

Sada se javlja problem alokacije prostora za Unicode poruku na medijumu koji se koristi. Ako je reč o nekom dokumentu na disku, on će da zauzima duplo više prostora nego konvencionalan dokument jer će se svaki znak zapisivati s dva bajta umesto samo jednim. Ako je reč o prenosu podataka preko računarske mreže, biće potrebno preneti duplo više podataka, pa će samim tim i prenos da traje duplo više (odnosno da košta duplo više). Postavlja se pitanje da li je to suviše velika cena za univerzalno pismo i da li postoji neki način da se taj problem prevaziđe i izbegne. Kao rešenje uvek stoji mogućnost da se zapisuje nekom odgovarajućom kodnom stranicom i troši bajt po znaku, ako nije neophodno korišćenje više pisama u istom dokumentu (što se retko dešava). Drugo rešenje je korišćenje tzv transformacionih šeme za pogodniji zapis i prenos podataka korišćenjem Unicode-a.

Prvo je razvijena Unicode transformaciona šema sa osnovnom jedinicom od 8 bita (UTF-8[2]). Pomoću nje se znak zapisuje u jednom, dva ili tri bajta, u zavisnosti od toga o kom je znaku reč. Ova transformaciona šema je prevashodno zgodna za upotrebu u jezicima koji koriste latinicu. O UTF-8 će biti više reči u poglavlju 3.

Jedan deo Mail Transfer Agent-a (MTA, program koji služi za prenošenje elektronske pošte na mail server-u), kao i zvanični standard za Internet Mail (IETF: STD 11, RFC 822) podržava samo 7-bitne mail poruke. MIME[3] standardi (RFC 2045 do 2049 [11], [12], [8], [9] i [10]) omogućavaju prenos višebitnih reči preko Internet mail-a, koristeći Base64[4] i Quoted Printable[5] načine kodiranja, međutim, oni nisu pravljeni za prenos Unicode-a nego za prenos bilo kakvih fajlova i nisu bili najoptimalnija rešenja. Zbog toga je kasnije razvijena 7-bitna transformaciona šema UTF-7. Tu se znak zapisuje u jednom ili u nekoliko bajtova, slično kao i u UTF-8. Osnovna razlika je u tome što UTF-7 koristi samo znakove Base64 koji bez problema mogu da se prenose putem elektronske pošte. Za takvu namenu se pokazalo da je UTF-7 optimalniji zapis nego UTF-8 kada se kodira sa Base64 ili sa Quoted Printable algoritmima kodiranja.

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, pošto još uvek ni jedan znak nije alociran u novodobijeni prostor, odnosno svi za sada definisani znaci (ceo UCS-2) se nalaze u ravni 0 ili osnovnoj višejezičkoj ravni (Basic Multilingual Plane, BMP). Međutim, pošto je UCS-4 novi standard za Unicode, treba i njega imati u vidu. Da bi se UCS-4 transparentno uveo u upotrebu redefinisani su formati zapisa UTF-7, UTF-8, UTF-16 i UTF-32. To je učinjeno tako da svaki znak iz UCS-2 ima istu reprezentaciju u UTF-7 i UTF-8 kao i ranije. UTF-16 je u neku ruku sinonim za UCS-2 i sadrži više od dva bajta samo u slučaju da se kodira neki znak van „Osnovne jezičke ravni" (BMP), koji za sada

---

[2] UTF = Unicode Transformation Format, odnosno „Oblik izmene Unicode-a"

[3] MIME = Multipurpose Internet Mail Extensions, tj. Višenamenska proširenja Internet pošte

[4] Base64 – 7bitni zapis koji koristi samo mala i velika slova latinice i karaktere +, / i =. Za više informacija pogledajte RFC 2045 [11]

[5] Quoted Printable je način kodiranja pomoći karaktera ASCII seta koji mogu da se odštampaju, tj. bez kontrolnih ASCII karaktera. Karakteri koji ne pripadaju tom setu se prikazuju u obliku **=<hex>** gde **<hex>** predstavlja heksadecimalnu vrednost koda karaktera. Za više informacija pogledajte RFC 2045 [11]

ne postoje. Za više informacija, pogledajte tabelu 1. UTF-32 je u stvari način zapisa UCS-4 u kome se koriste sva četiri bajta. Zbog toga što viši i niži bajt (ili dva bajta) mogu da se zapišu u memoriju na dva načina, postoje još po dve podvarijante UTF-16 i UTF-32 koje se razlikuju po redosledu bajtova. To su UTF-16BE (big-endian)[6] i UTF-16LE (little-endian)[7] i UTF-32BE i UTF-32LE. Ovo nije uvedeno da bi se uvela dodatna zabuna i zbrka, nego zato što različite arhitekture računara različito čuvaju podatke.

Takođe bih želeo da napomenem da postoje dve organizacije koje definišu dva standarda za Unicode. Jedan format je razvijen od strane tzv. The Unicode Consortium[8] pod nazivom The Unicode Standard [2]. Drugi standard je razvila Međunarodna organizacija za standardizaciju - International Standardization Organization, ISO[9] pod nazivom ISO/IEC 10646. Ta dva standarda su skoro identična i razlikuju se po pitanju tzv. Han unifikacije (predstavljanje Japanskih, Kineskih i Korejskih znakova jednim jedinstvenim skupom znakova), oko dodatnih znakova za definisanje akcenata, a od skoro i u tome što Unicode Consortium nije još podržao UCS-4 standard. Međutim, za našu upotrebu slobodno možemo da smatramo da su potpuno identični. Međunarodna organizacija koja definiše standarde za Internet - Internet Engineering Task Force, IETF[10] je u svojim standardima, tzv. „zahtevima za komentarima" (Request for Comments, RFC), u kojima je definisano sve što postoji na Internetu, prihvatila UTF-7 (RFC1642 [4] i RFC2152 [5]), UTF-8 (RFC2044 [14] i RFC2279 [15]) i UTF-16 (RFC2781 [13]), čime su oni i „zvanično" ušli u upotrebu na Internetu, tj. svuda. U najnovijim standardima IETF je izostavio Unicode Consortium i koristi samo verziju ISO 10646, što znači da je zvanično priznata verzija ISO 10646.

U HTML[11] jeziku za opis Web stranica se javljaju još dva načina za kodiranje Unicode znakova. Ovi načini traže mnogo više prostora nego originalni Unicode zapis i namenjeni su za korišćenje unutar neke od kodnih stranica za ubacivanje ponekog znakova iz neke druge kodne stranice. Jedan način je zapis oktalnih vrednosti UTF-8 bajtova. Zapisuje se tako što se prvo zapiše znak \, pa onda oktalna vrednost bajta. Ako taj znak u UTF-8 kodiranju sadrži više bajtova, svaki bajt se zapisuje na isti način. Tako na primer znak Φ čiji je UCS-2 kod `U+0424`[12], a UTF-8 zapis `0xD0 0xA4`[13] ima svoj HTML oktalni zapis kao `\320\244`, pošto je `0xD0 = 0320`[14] `= 208` i `0xA4 = 0244 = 164`.

Drugi način zapisa Unicode znakova u HTML-u je putem decimalne vrednosti njihovog UCS-2 koda. Zapisuje se tako što se prvo zapišu znakovi `&#`, pa onda decimalna vrednost UCS-2 koda i na kraju znak `;`. Tako bi se, na primer, gore pomenuti

---

[6] big-endian – kod **`U+abcd`** se zapisuje kao **`ab cd`**
[7] little-endian – kod **`U+abcd`** se zapisuje kao **`cd ab`**
[8] The Unicode Consortium – http://www.unicode.org/
[9] ISO – http://www.iso.org/
[10] IETF – http://www.ietf.org/
[11] HTML – Hyper-Text Markup Language
[12] Uobičajeno je da se naglasi da je reč o Unicode karakteru tako što se ispred heksadekadnog zapisa koda doda **`U+`**
[13] Heksadecimalan broj – Uobičajeno je da se heksadecimalan broj zapisuje sa prefiksom **`0x`** ili sa sufiksom **`h`**, da bi bilo jasno o čemu je reč. Ovaj standard je uzet iz programskog jezika C, koji je dominantan na UNIX operativnim sistemima, koji su dominantni u akademskom svetu koji definiše standarde :)
[14] Oktalni broj – Uobičajeno je da se oktalni broj zapisuje sa prefiksom **`0`** ili sa sufiksom **`o`**

znak Φ sa UCS-2 kodom `U+0424` zapisao u HTML decimalnom zapisu kao `&#1060;`, pošto je `0x0424 = 02044 = 1060`.

3. Ukratko o UTF-8

UTF-8 je zamišljen kao format koji najviše odgovara latiničnom tekstu. To je veoma pogodno za korišćenje u izvornom kodu programa ili u raznim jezicima ya markiranje (HTML, XML, LaTeX, ...) jer su standardne komande tih jezika uvek ASCII, a tekst koji se koristi može da bude i ASCII i UTF-8. Tako se ne ometa rad programskog kompilatora ili parsera jezika za markiranje, a omogućava se korišćenje višejezičke podrške.

U UTF-8 se znak zapisuje u obliku jednog bajta ako u svom zapisu sadrži samo najnižih 7 bita, odnosno, ako je reč o ASCII znaku (vidi odeljak 1). Ako znak u svom Unicode zapisu sadrži samo najnižih 11 bita, u UTF-8 se zapisuje u obliku dva bajta. I na kraju, ako znak sadrži svih 16 bita, zapisuje se u obliku tri bajta. U tabeli 1 je data šema kako se UCS-4 transformiše u UTF-8. Tabela je data za pun, četvorobajtni Unicode, a ako je reč o dvobajtnom Unicode-u, tj. o UCS-2, treba gledati samo prva tri reda u tabeli. Detaljniji opis algoritma za transformaciju može da se nađe u RFC2279 [15].

UTF-8 nije najoptimalniji način zapisa za kineski i japanski tekst, jer umesto da se koriste dva bajta po znaku, za takav tekst bi bilo korišćeno čak tri bajta po znaku, ali to i nije toliko važno za nas. Za ćirilični tekst je, s druge strane, svejedno da li se koristi čisti Unicode ili UTF-8, pošto se svaki ćirilični znak 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 znakovi zapisuju samo jednim bajtom (osim šđčćž).

Tabela: Šema kodiranja UCS-4 u UTF-8

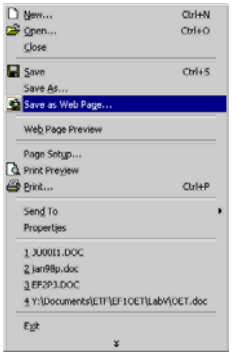
UCS-4 opseg (hex.)	UTF-8 binarni zapis
`0000 0000-0000 007F`	0xxxxxxx
`0000 0080-0000 07FF`	110xxxxx 10xxxxxx
`0000 0800-0000 FFFF`	1110xxxx 10xxxxxx 10xxxxxx
`0001 0000-001F FFFF`	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx
`0020 0000-03FF FFFF`	111110xx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx
`0400 0000-7FFF FFFF`	1111110x 10xxxxxx ...10xxxxxx

4 Korišćenje UTF-8 u programima

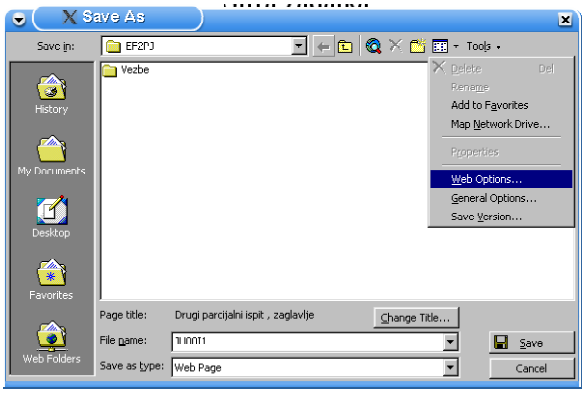
Microsoft Word for Windows. Jedan od nažalost najčešće korišćenih programa za obradu teksta pod,,operativnim sistemom" Windows jeste Microsoft Word for

Windows. On u svom formatu već ima podršku za više jezika. Unicode koristi u verziji 2002 (odnosno u OfficeXP-u), a za ranije verzije nisam siguran. Ako je potrebno da se taj Word dokument prebaci na Internet u obliku HTML fajla, potrebno je naglasiti da se sačuva u UTF-8 formatu. To otprilike izgleda ovako:

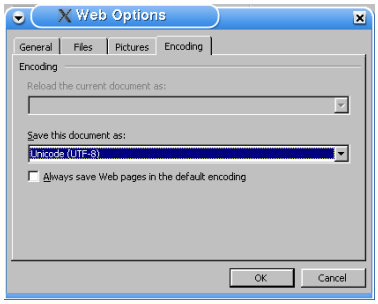
1. Prvo treba napisati sam dokument, naravno :)
2. Kad se dokument prebacuje na Internet, treba iz *File* menija izabrati opciju *Save as Web* (vidi sliku 1).
3. U dijalogu koji će da se pojavi treba izabrati gde se čuva fajl i pre nego što se stvarno sačuva, treba iz menija *tools* izabrati opciju *Web Options* (vidi sliku 2).
4. U novootvorenom dijalogu treba izabrati stranicu *Encoding* i tu u polju *Save this document as* izabrati *Unicode (UTF-8)* (vidi sliku 3).



Slika**:** WinWord - Sačuvaj kao Web stranicu



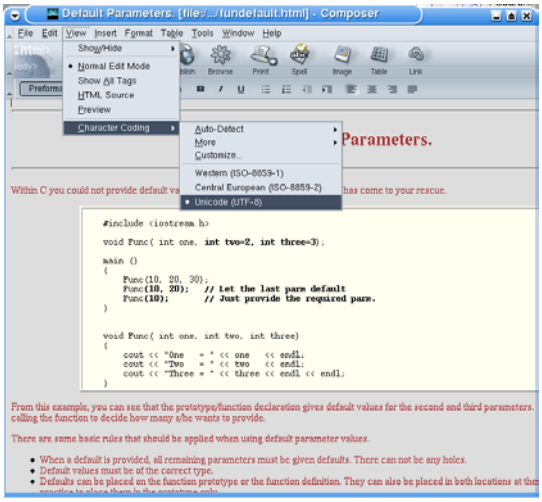
Slika: WinWord - Sačuvaj kao Web stranicu - dijalog



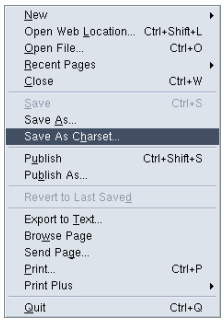
Slika 3: WinWord - Biranje UTF-8 koriranja

Netscape Communicator – Composer. Popularni Web čitač Netscape Communicator u svom sklopu ima i editor za Web stranice, tzv. Composer. On naravno može da bira kako će da čuva Web stranice i može da izabere i Unicode i to i UTF-8 i UTF-7. Nas zanima samo UTF-8, mada je postupak i za druge manje-više isti. Postoje dva načina da se u Netscape Composer-u tekst sačuva u UTF-8 formatu. Moguće je jednostavno izabrati iz menija *View* opciju *Charset* i tu izabrati UTF-8 kao format (vidi sliku 4).Posle toga se fajl najnormalnije sačuva u UTF-8 formatu.

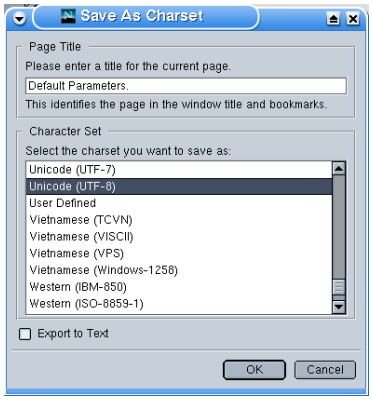
Drugi način je da se fajl sačuva umesto opcijom *File-Save,* opcijom *File-Save As Charset* (vidi sliku 5). Tada se dobija dijalog u kome može da se izabere način zapisa fajla i tu treba izabrati UTF-8 (vidi sliku 6).



Slika: Netscape Composer - Podesi način kodiranja

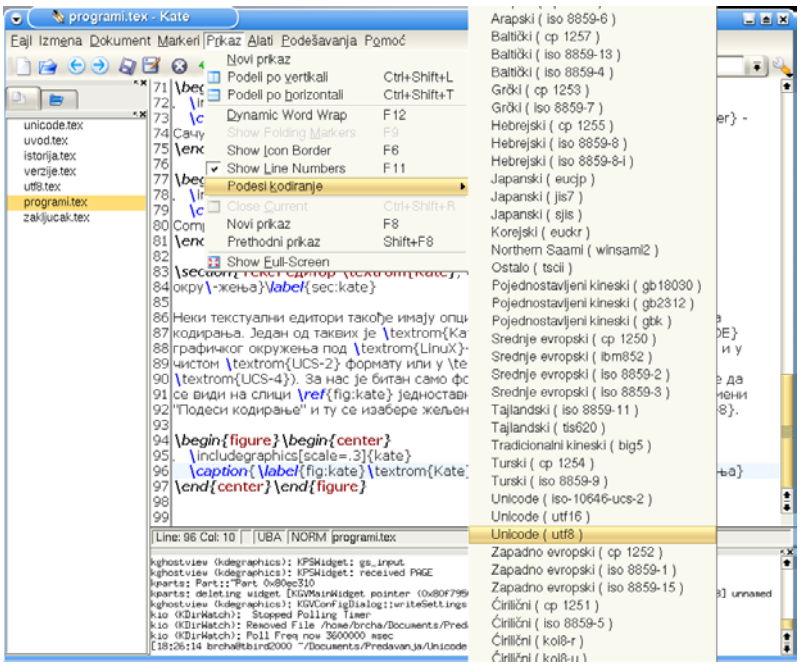


Slika: Netscape Composer - Sačuvaj sa specifičnim načinom kodiranja



Slika 6: Netscape Composer - Biranje UTF-8 kodiranja

Editor tekta Kate. Neki editori teksta takođe imaju mogućnost da tekst kodiraju na više načina. Jedan od takvih je Kate, koji je sastavni deo KDE grafičkog okruženja pod operativnim sistemom Linux. On takođe može da sačuva tekst i u čistom UCS-2 formatu ili u UTF-16 (kao delu UCS-4). Za nas je bitan samo format UTF-8. Kao što može da se vidi na slici 7 jednostavno se iz menija *Prikaz* izabere podmeni *Podesi kodiranje* i tu se izabere željeno kodiranje, odnosno UTF-8.



Slika 7: Kate - Biranje UTF-8 kodiranja

Iconv - konvertor kodiranja. Na operativnim sistemima Unix postoji biblioteka *iconv*[15] koja veoma jednostavno konvertuje iz jednog u drugi način kodiranja Za više informacija pogledajte [6]. Postoji i ekvivalentan command line program koji konvertuje fajlove iz Bilo kog načina kdiranja u bilo koji drugi način. Postoji lista kodova iz kojih i u koje ova biblioteka/program može da konvertuje: (pošto ovaj spisak zauzima više od 3 strane, izostavio sam ga, ali se može reći da podržava sve moguće načine kodiranja!).

Fontovi koji podržavaju Unicode. Da bi se koristio Unicode u pripremi dokumenata, potrebno je imati odgovarajuće fontove koji ga (barem delimično) podržavaju. Od fontova dostupnih na Windows-u, Unicode sigurno podržavaju fontovi Arial, Times New Roman, Helvetica, Verdana i Courier New, a takođe su instalirani na svim Windows-platformama, tako da bi generalno trebalo da se koristi neki od tih fontova. Fontovi tipa TimesCirilica ili YULTimes mogu da prikažu naša slova, ali su daleko od Unicode-a i u prenosu fajla sa jednog na drugi računar u elektronskom obliku postoji velika šansa da taj fajl neće biti čitljiv na drugom računaru, tako da bi trebalo da se takvi nestandardni fontovi izbegavaju koliko god je to moguće.

Na Linux-u i ostalim Unix-ima se u samom nazivu fonta vidi da li podržava unicode ili ne, pošto poslednji deo naziva fonta predstavlja *character set* fonta. Ako tu piše iso10646, to znači da je font kompatibilan s Unicode-om. Međutim, i ovde bih preporučio, radi prenosivosti dokumenata, da se koriste standardni (Adobe) fontovi, kao što su Times (-adobe-times-*-iso10646-1), Utopia (-adobe-utopia-*-iso10646-1), Helvetica (-adobe-helvetica-*-iso10646-1), Courier (-adobe-courier-*-iso10646-1).

5 Unicode, baze podataka i XML

Svetski trendovi razvoja baza podataka idu ka uvođenju Unicode-a, kao standardnog načina zapisa podataka i XML-a[16], kao standarnog jezika za prenos i prezentaciju tih podataka.

Većina baza podataka već duže vreme podržava Unicode. Dobar deo aplikacija za rad s bazama koristi XML za prezentovanje i prenos podataka, zato što se pokazalo da je XML jednostavan jezik za programiranje, za koji već postoji puno parsera[17], i zato što se pokazalo da je on dovoljno fleksibilan da može da prenese bilo kakav tip podataka na sličan način. Da bi se programi međusobno ,,razumeli" razvijeni su razni standardi za opis podataka u XML-u (kao što je na primer Encoded Archival Description Standard [7]).

---

[15] IConv – *i*nternational *conv*ersion

[16] XML - eXtendable Markup Language (,,Proširivi jezik za obeležavanje"), naslednik SGML-a (Standard Generalized Markup Language, tj. ,,Standardni generalizovani jezik za obeležavanje") - fleksibilan jezik za zapisivanje proizvoljnih podataka na standardni način, što ga čini vrlo pogodnim za prezentovanje podataka iz baze podataka. Za više informacija pogledajte [3] i [1]

[17] Parser je program koji analizira neki sadržaj tako da ga razdvoji na sastavne činioce, kako bi se oni dalje obrađivali. Recimo da u nekom XML fajlu postoji sledeća sekvenca ,,<osoba> <ime>Petar</ime> <prezime>Petrović</prezime> <email>petar.petrovic@transmeta.com</email> </osoba>". Iz ove, za program nerazumljive sekvence znakova, odgovarajući parser će da izdvoji ime (Petar), prezime (Petrović) i email(petar.petrovic@transmeta.com) i posle će sa time program dalje moći da nešto radi (npr. da pošalje e-poruku Petru Petroviću)

To uvođenje XML-a kao glavnog jezika za podršku bazama podataka je još više učvrstilo poziciju Unicode-a, pošto se XML fajlovi standardno pišu u UTF-8 ili UTF-16. Zanimljivo je da je Microsoft, koji se uglavnom protivi svim standardima i trudi se da definiše svoje, prihvatio XML i koristi ga gde god može. Cela .NET tehnologija je bazirana na XML-u. Zbog toga može da se očekuje da će u budućnosti biti samo više XML-a i više Unicode-a i da je bitno što ranije se orjentisati ka njima.

6. Tabela kodova za naša slova

U tabeli 2 su data skoro sva slova koja se kod nas koriste, s odgovarajućim UCS-2 kodom, UTF-8 zapisom i sa HTML oktalnim i decimalnim zapisima (za više informacija pogledajte odeljak 2).

Slovo	Izgled	UTF-8	oktalna	decimalna
Velika latinična slova				
U+0041	A	0x41	\101	A
U+0042	B	0x42	\102	B
U+0043	C	0x43	\103	C
U+0044	D	0x44	\104	D
U+0045	E	0x45	\105	E
U+0046	F	0x46	\106	F
U+0047	G	0x47	\107	G
U+0048	H	0x48	\110	H
U+0049	I	0x49	\111	I
U+004A	J	0x4A	\112	J
U+004B	K	0x4B	\113	K
U+004C	L	0x4C	\114	L
U+004D	M	0x4D	\115	M
U+004E	N	0x4E	\116	N
U+004F	O	0x4F	\117	O
U+0050	P	0x50	\120	P
U+0051	Q	0x51	\121	Q
U+0052	R	0x52	\122	R
U+0053	S	0x53	\123	S
U+0054	T	0x54	\124	T
U+0055	U	0x55	\125	U
U+0056	V	0x56	\126	V
U+0057	W	0x57	\127	W
U+0058	X	0x58	\130	X

U+0059	Y	0x59	\131	Y
U+005A	Z	0x5A	\132	Z
Mala latinična slova				
U+0061	a	0x61	\141	a
U+0062	b	0x62	\142	b
U+0063	c	0x63	\143	c
U+0064	d	0x64	\144	d
U+0065	e	0x65	\145	e
U+0066	f	0x66	\146	e
U+0067	g	0x67	\147	g
U+0068	h	0x68	\150	h
U+0069	i	0x69	\151	i
U+006A	j	0x6A	\152	j
U+006B	k	0x6B	\153	k
U+006C	l	0x6C	\154	l
U+006D	m	0x6D	\155	m
U+006E	n	0x6E	\156	n
U+006F	o	0x6F	\157	o
U+0070	p	0x70	\160	p
U+0071	q	0x71	\161	q
U+0072	r	0x72	\162	r
U+0073	s	0x73	\163	s
U+0074	t	0x74	\164	t
U+0075	u	0x75	\165	u
U+0076	v	0x76	\166	v
U+0077	w	0x77	\167	w
U+0078	x	0x78	\170	x
U+0079	y	0x79	\171	y
U+007A	z	0x7A	\172	z
Naša dodatna latinična slova				
U+0106	Ć	0xC4 0x86	\304\206	Ć
U+0107	ć	0xC4 0x87	\304\207	ć
U+010C	Č	0xC4 0x8C	\304\214	Č
U+010D	č	0xC4 0x8D	\304\215	č
U+0110	Đ	0xC4 0x90	\304\220	Đ
U+0111	đ	0xC4 0x91	\304\221	đ

U+0160	Š	0xC5 0xA0	\305\240	Š
U+0161	š	0xC5 0xA1	\305\241	š
U+017D	Ž	0xC5 0xBD	\305\275	Ž
U+017E	ž	0xC5 0xBE	\305\276	ž
Velika ćirilična slova				
U+0402	Ђ	0xD0 0x82	\320\202	Ђ
U+0408	Ј	0xD0 0x88	\320\210	Ј
U+0409	Љ	0xD0 0x89	\320\211	Љ
U+040A	Њ	0xD0 0x8A	\320\212	Њ
U+040B	Ћ	0xD0 0x8B	\320\213	Ћ
U+040F	Џ	0xD0 0x8F	\320\217	Џ
U+0410	А	0xD0 0x90	\320\220	А
U+0411	Б	0xD0 0x91	\320\221	Б
U+0412	В	0xD0 0x92	\320\222	В
U+0413	Г	0xD0 0x93	\320\223	Г
U+0414	Д	0xD0 0x94	\320\224	Д
U+0415	Е	0xD0 0x95	\320\225	Е
U+0416	Ж	0xD0 0x96	\320\226	Ж
U+0417	З	0xD0 0x97	\320\227	З
U+0418	И	0xD0 0x98	\320\230	И
U+041A	К	0xD0 0x9A	\320\232	К
U+041B	Л	0xD0 0x9B	\320\233	Л
U+041C	М	0xD0 0x9C	\320\234	М
U+041D	Н	0xD0 0x9D	\320\235	Н
U+041E	О	0xD0 0x9E	\320\236	О
U+041F	П	0xD0 0x9F	\320\237	П
U+0420	Р	0xD0 0xA0	\320\240	Р
U+0421	С	0xD0 0xA1	\320\241	С
U+0422	Т	0xD0 0xA2	\320\242	Т
U+0423	У	0xD0 0xA3	\320\243	У
U+0424	Ф	0xD0 0xA4	\320\244	Ф
U+0425	Х	0xD0 0xA5	\320\245	Х
U+0426	Ц	0xD0 0xA6	\320\246	Ц
U+0427	Ч	0xD0 0xA7	\320\247	Ч
U+0428	Ш	0xD0 0xA8	\320\250	Ш
Mala ćirilična slova				

U+0430	а	0xD0 0xB0	\320\260	а
U+0431	б	0xD0 0xB1	\320\261	б
U+0432	в	0xD0 0xB2	\320\262	в
U+0433	г	0xD0 0xB3	\320\263	г
U+0434	д	0xD0 0xB4	\320\264	д
U+0435	е	0xD0 0xB5	\320\265	е
U+0436	ж	0xD0 0xB6	\320\266	ж
U+0437	з	0xD0 0xB7	\320\267	з
U+0438	и	0xD0 0xB8	\320\270	и
U+043A	к	0xD0 0xBA	\320\272	к
U+043B	л	0xD0 0xBB	\320\273	л
U+043C	м	0xD0 0xBC	\320\274	м
U+043D	н	0xD0 0xBD	\320\275	н
U+043E	о	0xD0 0xBE	\320\276	о
U+043F	п	0xD0 0xBF	\320\277	п
U+0440	р	0xD0 0xC0	\320\280	р
U+0441	с	0xD0 0xC1	\320\281	с
U+0442	т	0xD0 0xC2	\320\282	т
U+0443	у	0xD0 0xC3	\320\283	у
U+0444	ф	0xD0 0xC4	\320\284	ф
U+0445	х	0xD0 0xC5	\320\285	х
U+0446	ц	0xD0 0xC6	\320\286	ц
U+0447	ч	0xD0 0xC7	\320\287	ч
U+0448	ш	0xD0 0xC8	\320\290	ш
U+0452	ђ	0xD1 0x92	\321\222	ђ
U+0458	ј	0xD1 0x98	\321\230	ј
U+0459	љ	0xD1 0x99	\321\231	љ
U+045A	њ	0xD1 0x9A	\321\232	њ
U+045B	ћ	0xD1 0x9B	\321\233	ћ
U+045F	џ	0xD1 0x9F	\321\237	џ

Bibliografija

1. Filip Brčić, *Ukratko o XML-u*, http://brcha.free.fr/documents/XMLtut/xmltut.pdf.
2. The Unicode Consortium, *The Unicode Standard - Version 3.0,* Addison-Wesley, 2000 http://www.unicode.org
3. World Wide Web Consortium, *Extensible markup language (xml) 1.1, Candidate recommendation*, 2002 http://www.w3c.org/TR/xml11

4. M. Davis, D. Goldsmith, *Utf-7 – a mail-safe transformation format of unicode,* Experimental 1642, Internet Engineering Task Force, 1994, http://www.ietf.org/rfc/rfc1642.txt
5. M. Davis, D. Goldsmith, *Utf-7 – a mail-safe transformation format of unicode,* Informational 2152, Internet Engineering Task Force, 1997, http://www.ietf.org/rfc/rfc2152.txt
6. Ulrich Drepper, *MANPAGE: Iconv(3) 2.2.5 – Perform character set conversion*, Free Software Foundation, 2002.
7. Bojan Marinković, Encoded archival description document type definition, 2003.
8. K. Moore, *Multipurpose internet mail extensions (mime) part three: Message header extensions for non-ascii text,* Standards Track 2047, Internet Engineering Task Force, 1996, http://www.ietf.org/rfc/rfc2047.txt
9. J. Postel, N. Freed, J. Klensin, *Multipurpose internet mail extensions (mime) part four: Registration procedures,* Standards Track 2048, Internet Engineering Task Force, 1996, http://www.ietf.org/rfc/rfc2048.txt
10. N. Borenstein, N. Freed, *Multipurpose internet mail extensions (mime) part five: Conformance criteria and examples,* Standards Track 2049, Internet Engineering Task Force, 1996, http://www.ietf.org/rfc/rfc2049.txt
11. N. Borenstein, N. Freed, *Multipurpose internet mail extensions (mime) part one: Format of internet message bodies,* Standards Track 2045, Internet Engineering Task Force, 1996, http://www.ietf.org/rfc/rfc2045.txt
12. N. Borenstein, N. Freed, *Multipurpose internet mail extensions (mime) part two: Media types,* Standards Track 2046, Internet Engineering Task Force, 1996, http://www.ietf.org/rfc/rfc2046.txt
13. F. Yergeau, P. Hoffman,*Utf-16, an encoding of iso 10646,* Informational 2781, Internet Engineering Task Force, 2000, http://www.ietf.org/rfc/rfc2781.txt
14. F. Yergeau,
Utf-8, a trensformation format of unicode and iso 10646, Informational 2044, Internet Engineering Task Force, 1996, http://www.ietf.org/rfc/rfc2044.txt
15. F. Yergeau, *Utf-8, a trensformation format of iso 10646,* Standards Track 2279, Internet Engineering Task Force, 1998, http://www.ietf.org/rfc/rfc2279.txt

brcha@users.sourceforge.net

We review coding of multi-language text in digital form using Unicode standard, with special attention to UTF-8 variant, which is the most convenient variant for coding latin text. We also give a short tutorial for using UTF-8 in Microsoft Word, Netscape Composer and text editor Kate. Standard Unicode fonts are recommended so that the texts can be easily transfered from a computer to another one or for publishing on Internet.