Univerza v Ljubljani
Naravoslovnotehniška fakulteta
Oddelek za tekstilstvo
Grafična tehnologija

Seminarska naloga pri predmetu Tipografija v digitalnih medijih

PREDNOSTI IN POMANJKLJIVOSTI UNICODE PISAV

Maša Žveglič

Ljubljana, maj 2006

KAZALO

1 UVOD

Odkar je človek začutil potrebo, da bi izrazil svoja sporočila ali misli v pisani obliki, je v skladu z danimi možnostmi neprestano razvijal tudi tehniko in tehnologijo njihovega posredovanja. V današnjem svetu je to tehnologija računalništva, ki nam omogoča pisanje in oblikovanje črk. Črke in drugi znaki so v računalniku predstavljeni s pomočjo tabele, imenovane kodna tabela ali kodni razpored, ki povezuje grafično predstavitev nekega znaka z njegovim binarnim zapisom. Kot je v računalništvu navada, moramo biti pazljivi in se prepričati, v katerem številskem sistemu je navedeno število: razen desetiškega sistema se uporabljata se osmiški in šestnajstiški ter v izpisih redko dvojiški. Tako znaku za majuskulo A pri nekaterih kodnih tabelah ustreza desetiško število 65.

Unicode je standard za kodiranje znakov v računalništvu, ki glede na implementacijo za zapis znaka uporablja od enega do šestnajst bajtov. To naj bi zadoščalo za zapis večine svetovnih jezikov vključno z japonščino in s kitajščino. Razvit je bil pod okriljem organizacije Unicode Consortium, ki je organizacija, zadolžena za razvoj in koordinacijo danega standarda. V njej so prisotna tako rekoč vsa pomembnejša podjetja iz panoge računalniških in informacijskih tehnologij [1].

Za splet sta Microsoft in Adobe pred časom pripravila pisave OpenType, ki temeljijo na kodi Unicode in podpirajo najpogostejše abecede. Celovitejšo podporo kode Unicode dajejo pisave lucida sans unicode (na NT operacijskem sistemu), arial unicode (na XP operacijskem sistemu) in pisava cyberbit.

2 RAZVOJ UNICODE STANDARDA

V osnovi računalniki računajo samo s števili. Črke in ostale znake zapišejo kot določena števila. Preden so razvili Unicode, je obstajalo na stotine različnih kodnih tabel za dodeljevanje teh števil. Do leta 1964 so proizvajalci uporabljali 6-bitno BCD kodno tabelo, ki je vključevala 26 majuskul angleške abecede, 10 številk in 28 posebnih znakov. Ker to ni bilo dovolj, se je začela uporabljati 7-bitna ASCII kodna tabela, nato pa je nastala še 8-bitna ASCII kodna tabela, ki je dala 128 dodatnih znakov.

Nobena kodna tabela pa ni mogla vsebovati dovolj znakov. Na primer: Evropska Unija bi potrebovala kar nekaj različnih kodnih tabel, da bi podprla vse evropske jezike. Celo pri samo enem jeziku, kot je angleščina, ena kodna tabela ni primerna za vse črke, ločila in tehnične simbole. Prav tako so te kodne tabele v sporu ena z drugo. Kar pomeni, da dve kodni tabeli lahko uporabljata enako število za različne znake oziroma različna števila za enak znak. Vsi računalniki (posebej strežniki) morajo podpirati veliko različnih kodnih tabel. Kadar se podatki prenašajo preko različnih kodnih tabel ali platform, vedno obstaja možnost, da pride do napačnih podatkov.

Z uporabo Unicode kodne tabele se zgoraj opisane težave rešijo. Unicode uporablja edinstvena števila za vsak znak, ne glede na platformo, programsko opremo in jezik. Je 16-bitna kodna tabela, ki zajema 65.536 znakov, kar je dovolj za vse pisave na svetu.

Unicode standard so sprejela vodilna računalniška podjetja kot so Apple, HP, IBM, JustSystem, Microsoft, Oracle, SAP, Sun, Sybase, Unisys in drugi. Unicode se uporablja v novejših standardih, kot so XML, Java, ECMAScript, LDAP, COBRA 3.0, WM, in je uradno potreben za izvajanje standarda ISO/IEC 10646. Podpirajo ga mnogi operacijski sistemi, vsi novejši internet brskalniki ter številni drugi produkti. Prisotnost standarda Unicode in dostopnost orodij, ki ga podpirajo, je trenutno zelo pomembna smer na področju programske tehnologije.

Vključevanje standarda Unicode v sisteme odjemalec-strežnik ali večplastne aplikacije in spletne strani omogoča znaten prihranek glede na uporabo starih znakovnih naborov. Unicode omogoča uporabo posameznega programa ali posamezne spletne strani na različnih platformah, jezikih in državah, brez dodatnih tehničnih posegov. Omogoča prenos podatkov skozi različne sisteme brez napak ali napačnih podatkov [2].

2.1 Unicode konzorcij

Unicode konzorcij je neprofitna organizacija, ki je bila ustanovljena za razvoj, širjenje in promocijo standarda Unicode, ki določa prikaz teksta v novejših programskih izdelkih in standardih. Članstvo v konzorciju predstavlja širok spekter družb in organizacij v računalniški in informacijski industriji, finančno pa se oskrbuje izključno s članarino. Članstvo v konzorciju Unicode je odprto za organizacije in posameznike povsod po svetu, ki so pripravljeni pomagati standardu Unicode pri njegovem širjenju in izvajanju [2].

2.2 Standard ISO 10646

Leta 1991 sta organizacija ISO, ki je odgovorna za standard ISO/IEC 10646, in Unicode konzorcij ustvarila univerzalni standard za kodiranje znakov večjezičnega besedila. Čeprav so kodni znaki in nekodne oblike usklajene s strani Unicode in ISO/IEC 10646, Unicode standard omogoča uporabo edinstvenih števil za vsak znak, ne glede na platformo in programsko opremo, in nadgrajuje funkcionalne sposobnosti znakov, podatkovno bazo znakov in algoritme, ki niso zapisani v ISO/IEC 10646 [2].

V četrti izdaji ISO 10646:2003 UCS (Universal Character Set) je 31-bitni nabor znakov, razdeljen na 128 »skupin«, vsaka od njih pa na 256 »ravnin«; 65536 znakov prve ravnine je definiranih kot osnovna večjezična ravnina (BMP – Base Multilangual Plane). V tretji izdaji standarda Unicode (enakovredni standardu ISO/IEC 10646-1:2000) je definiranih 49.194 znakov, 7827 16-bitnih kod je še nedodeljenih, če štejemo še 1.048.544 za zdaj nedodeljenih mest, ki jih lahko naslavljamo s pari surogatov (torej s parom 16-bitnih kod), pa bi moral standard za nekaj časa zadostovati.

Za kodiranje besedil se ob kodiranju UCS-2, v katerem je vsak od 65.536 znakov predstavljen z dvema bajtoma, ter UCS-4, v katerem je vsak znak predstavljen s štirimi bajti, uporablja tudi pretvorna shema UTF-8 (UCS Transformation Scheme), v kateri se prvih 128 znakov iz nabora Unicode (kar ustreza naboru ASCII) kodira z enim bajtom, naslednjih 1920 (med njimi je večina drugih latiničnih znakov, cirilica, osnovna grščina, hebrejščina in osnovna arabščina) kot dva bajta, nadaljnjih 63.488 znakov (med njimi kitajski, japonski, korejski) s tremi, preostali znaki iz nabora ISO 10646 pa s štirimi do šestimi bajti. Prednost kodiranja UTF-8 je za evropske latinično pišoče narode v tem, da so besedila komaj kaj daljša od besedil, kodiranih v katerem od osembitnih standardov ISO 8859, kodiranje v UCS-2 pa pomeni dvakrat daljše besedilo (UCS-4 pa celo štirikrat daljše). Slabost je seveda v tem, da v takih besedilih ni več enostavne zveze med številom znakov v besedilu in dolžino besedila v bajtih [2].

3 UPORABA UNICODE

Ljudje v različnih deželah zapisujejo svoj materni jezik z različnimi črkopisi. Dandanes je večina aplikacij, skupaj s sistemi za elektronsko pošto in spletnimi brskalniki, 8-bitna, kar pomeni, da lahko pravilno prikažejo besedilo, če je napisano v enem izmed 8-bitnih naborov znakov, na primer ISO-8859-1 ali ISO-8859-2.

Po svetu je v rabi precej več kot 256 znakov, kar se kaže predvsem pri cirilici, hebrejščini, arabščino, kitajščini, japonščini, korejščini in tajščini, še vedno pa se uvajajo kakšni novi znaki. Prav zaradi tega lahko uporabnik naleti na naslednje težave:

- Nemogoče je shraniti besedilo z znaki iz različnih naborov znakov. Na primer v publikaciji v nemščini, francoščini ali slovenščini je mogoče citirati članek v ruščini, če uporabnik uporablja TeX, Xdvi in PostScript, ne more pa tega storiti v čistem besedilu.
- Dokler je vsak dokument napisan v svojem naboru znakov in ta nabor ni prepoznan avtomatsko, so ročne nastavitve neizogibne.
- Uvajajo se novi simboli (npr. za evro), zato se je uvedel tudi nov standard ISO-8859-15, ki se večinoma ujema z ISO-8859-1, le da so odstranjeni nekateri redko rabljeni znaki (stari znaki za valute) in je dodan znak za evro. Če uporabniki sprejmejo ta standard, imajo na disku dokumente v različnih naborih znakov in se morajo vsak dan ubadati s tem. Računalniki pa bi morali stvari poenostaviti, ne še bolj zaplesti.

Rešitev teh težav je sprejetje po vsem svetu uporabnega nabora znakov. Tak nabor znakov je Unicode, ki je standard za 16-bitno kodno tabelo, katera lahko vsebuje znake vseh pisav vsega sveta [3].

Primer prepoznavanja znakov arabske pisave:

a) V prvem primeru je stavek napisan s pisavo gramond, katera ne prepozna arabskih znakov.

Z arabsko pisavo se to zapiše □□□□□.

b) V drugem primeru je stavek napisan s pisavo lucida sans unicode, katera prepozna arabske znake.

Z arabsko pisavo se to zapiše □□□□□.

3.1 Kodiranja po Unicode

Unicode resda lahko odpravi probleme različnih kodnih strani, prinese pa tehnično težavo, in sicer na kakšen način zapisati znake Unicode z 8-bitnimi zlogi. 8-bitni zlog je pri večini računalnikov najmanjša naslovljiva enota in tudi osnovna enota pri omrežnih povezavah prek protokola TCP/IP. Uporaba enega zloga za predstavitev enega znaka je zgodovinsko naključje, predvsem posledica dejstva, da se je razvoj računalništva začel v Evropi in ZDA, kjer je 96 znakov zadostovalo za dolgo vrsto let.

V osnovi poznamo štiri načine kodiranja znakov Unicode v zloge:

- **UTF-8**: 128 znakov se kodira z uporabo enega zloga (znaki ASCII). 1920 znakov se kodira z uporabo dveh zlogov (rimski, grški, cirilični, koptski, armenski, hebrejski in arabski znaki). 63488 znakov se kodira z uporabo treh znakov (med drugim kitajski in japonski znaki). Preostalih 2147418112 znakov (ki še niso povsem določeni) se lahko kodira z uporabo 4, 5 ali 6 zlogov.
- **UCS-2**: Vsak znak je predstavljen z dvema zlogoma. Tako kodiranje lahko predstavi le prvih 65536 znakov iz Unicode.
- **UTF-16**: To je razširitev UCS-2, ki lahko predstavi 1112064 znakov iz Unicode. Prvih 65536 znakov je predstavljenih z dvema zlogoma, drugi s štirimi.
- **UCS-4**: Vsak znak je predstavljen s štirimi zlogi [3].

Prostorske zahteve za kodiranje besedil v primerjavi s trenutno rabljenimi (8-bitni za evropske jezike, več za kitajščino,japonščino ali korejščino) so razvidne iz spodnjega opisa. Pri tem gre za porabo prostora na disku in hitrost prenašanja po omrežju, če ni uporabljena nobena oblika stiskanja:

- **UTF-8**: Nobene spremembe za US ASCII, samo nekaj odstotkov več za ISO-8859-1, 50 % več za kitajske, japonske ali korejske znake, 100 % več za grške in cirilične znake.
- **UCS-2 in UTF-16**: Nobene spremembe za kitajske, japonske ali korejske znake. 100 % več za US ASCII, ISO-8859-1, ISO-8859-2, grške in cirilične znake.
- **UCS-4**: 100 % več za kitajske, japonske ali korejske znake. 300 % več za US ASCII, ISO-8859-1, ISO-8859-2, grške in cirilične znake.

Prihodnost v razvoju pripisujejo UTH-8, ker zajema številne jezike in je brez dodatnih zahtev prenosljiv v programe za urejanje besedila 3].

Posebni programi, ki temeljijo na naboru znakov iz Unicode kodne tabele so:

- programi za delo z omrežjem (telnet in kermit),
- brskalniki (Netscape, Mozilla, Lynx, W3M),
- urejevalniki (Yudit, Vim, Emacs, Nedit, Xedit, Axe, Pico, Mined98),
- programi za elektronsko pošto (Pine, Netscape Communicator, Emacs, Mutt, Exmh)
- programi za obdelavo besedil (Groff, TeX),
- zbirke podatkov (PostgreSQL) in
- drugi programi v tekstovnem načinu (Less, Lv, Expand, Wc, Col, Colcrt, Colrm, Column, Rev, Ul, Figlet) [3].

3.2 Pisave truetype

V nekaterih aplikacijah, še posebej v tistih za tiskanje, so nujno potrebne pisave z veliko ločljivostjo. Najpomembnejša vrsta pisav z nastavljivo velikostjo in veliko ločljivostjo so pisave truetype.

Nekatere brezplačne pisave truetype, ki pokrivajo precejšen del Unicode, so:

- **bitstream cyberbit:** Pokriva cirilico, rimsko, grško, hebrejsko, arabsko, kitajsko, korejsko, japonsko in druge pisave skupaj s kombiniranimi diakritičnimi znamenji.
- **microsoft arial:** Pokriva cirilico, rimsko, grško, hebrejsko, arabsko, vietnamsko pisavo in nekatera kombinirana diakritična znamenja.
- **lucida sans unicode:** Pokriva cirilico, rimsko, grško, hebrejsko pisavo in kombinirana diakritična znamenja [3].

3.3 Tiskanje z uporabo pisav truetype

Ker postscript sam po sebi ne podpira pisav Unicode, morajo vso odgovornost za podporo Unicode pri tiskanju prevzeti programi, ki naredijo datoteko v postscriptu.

Tako uniprint (tekstovno datoteko pretvori v postscript) kot wprint – WorldPrint (naknadno obdela izhod v postscriptu) omogočata dobro tiskanje tekstovnih datotek, ki so kodirane po Unicode. Oba zahtevata nameščene pisave truetype. Pri tekstovnih datotekah je splošna razporeditev pri uniprintu nekoliko boljša, vendar pa samo wprint pravilno izpiše besedila v tajščini [3].

3.4 Uporaba Unicode pisav na spletu

Razširljivi označevalni jezik XML − eXstensible Markup Language je nastal kot posledica slabosti in pomanjkljivosti, ki jih imata njegova predhodnika: standardni splošni označevalni jezik SGML − Standard Generalized Markup Language in jezik za označevanje hiperteksta HTML − Hyper Text Markup Language. SGML in HTML jezika bazirata na ASCII − American standard code for information interchange standardu, XML pa bazira na Unicode sistemu, ki omogoča uporabo več znakov, kar pomeni, da lahko v XML-u uporabimo tudi nabore črk tujih abeced.

XML je programski jezik za opisovanje strukture in pomena podatkov. XML je zbirka pravil za definiranje pomenskih oznak. Za razliko od drugih označevalnih jezikov, kot je na primer jezik za označevanje hiperteksta HTML, ne vsebuje vnaprej definirane množice oznak, ampak omogoča določanje lastnih oznak, ki morajo biti v dokumentu urejene glede na ustrezna pravila [3].

4 PREDNOSTI UNICODE PISAV

Največja prednost Unicode pisav je v Unicode standardu, ki pokriva vse svetovne jezike, ki so v rabi danes. Poteka tudi že vključevanje zgodovinskih pisav (npr.: egipčanski hieroglifi), znakov glasbene notacije in zelo redkih kitajskih ideogramov.

Velika prednost Unicode pisav je tudi v tem, ker jo podpirajo vse zadnje verzije pomembnih operacijskih sistemih, vsi pomembnejši programski jeziki in zadnje verzije spletnih brskalnikov, saj je Unicode privzel nabor znakov v zadnjih verzijah HTML in XML jezika.

Nekateri jeziki (predvsem vzhodnoazijski, kot so japonščina, različna kitajska in korejska narečja) ne morejo biti predstavljeni samo z 256 znaki, ker gre za znake, ki predstavljajo posamezne besede (pismenke), katerih število različnih znakov ponavadi presega 6000. Unicode standard rešuje to težavo s 16-bitno kodno tabelo, ki zajema 65.536 znakov, kar je dovolj za vse pisave na svetu. Unicode standard omogoča svobodno mešanje zelo različnih pisav v istem dokumentu [4].

5 SLABOSTI UNICODE PISAV

Slabosti Unicode pisav je v tem, da še ne vsebuje vseh matematičnih formul in razne posebne znake. Prav tako Unicode znaki niso možni pri imenih makrojev, plasti, risarskih datotek, načrtov in projektov. Problemi se pojavijo tudi pri prikazu kompleksnih kombinacij, kot je prikazano na spodnjem primeru:

a + ˆ + ˚ = å

Unicode standardu mora biti prilagojena tudi vsa programska oprema, ki dela z dokumenti. Ker standardnega Unicodovega nabora znakov ne podpirajo vsi spletni strežniki, se lahko znaki uporabljajo samo v glavnih spletnih strežnikih. Vendar izvor tovrstne »Unicodove slabosti« ni v naboru znakov Unicode, temveč v načinu njegove uporabe na programski strani.

Trenutno obdelave kodiranja znakov po Unicode standardu zahtevajo veliko procesiranja (posebno pri UTF8) in veliko podatkovnega prostora. Vendar z razvojem računalniške zmogljivosti se bo ta težava v prihodnosti odpravila in uporaba Unicode kodiranja bo hitra ter preprosta [5].

6 ZAKLJUČEK

Unicode standard je poskus vzpostavitve enotnega mednarodnega standarda kodne razporeditve znakov vseh svetovnih jezikov. Je uporaba enotne oblike zapisa pisav za prikaz črk na zaslonu, za namizno založništvo, natis s tiskalniki in za splet, kar je v današnji globalizirani družbi še kako pomembno.

Zaradi svoje specifičnosti (ligature) so kot poseben izziv kažejo indijske pisave, ki se jim bo zato v prihodnjih letih verjetno namenilo največ pozornosti. Iz več razlogov pa so razmeroma nizko na prioritetni lestvici vzhodnoazijske ideografske pisave, ker se standard ISO 10646/Unicode zaradi obstoječih lokalnih standardov v tem delu sveta počasi uveljavlja, in ne nazadnje, dodajanje nekaj deset tisoč pismenk poveča datoteke s pisavami do mere, ko z današnjimi računalniškimi zmogljivostmi delo postane neudobno. Zadnja težava bo z naraščanjem računalniških zmogljivosti verjetno v nekaj letih odpravljena, in tedaj se načrtuje tudi vključitev teh znakov.

7 LITERATURA IN VIRI

[1] *Unicode* [online]. Wikipedija, obnovljeno 2006 [citirano 23. 5. 2006]. Dostopno na svetovnem spletu: <http://sl.wikipedia.org/wiki/Unicode>.

[2] Unicode [online], 1991, obnovljeno 2006 [citirano 23. 5. 2006]. Dostopno na svetovnem spletu: <http://www.unicode.org>.

[3] Haible B., Unicode HOWTO [online], 2002 [citirano 23. 5. 2006]. Dostopno na svetovnem spletu: <http://old_www.lugos.si/delo/slo/HOWTO-sl/Unicode-HOWTO-sl.html>.

[4] *Center za uporabnike* [online]. Microsoft, 2006 [citirano 243. 5. 2006]. Dostopno na svetovnem spletu: <http://www.microsoft.com/slovenija/mscenter/nasveti/2001-02.mspx>.

[5] Bratuša T., Spletni brskalnik −Ubijalsko hekersko orodje, v: Monitor, 7/8, 2005, str. 36-40.