EVS JUHEND 8

See dokument on EVS-i poolt koostatud eelvaade

STANDARDITE ISO/IEC 10646 JA UNICODE KASUTUSJUHEND

Guidelines for using of ISO/IEC 10646 and UNICODE

EESTI STANDARDIKESKUS **2005** AMETLIK VÄLJAANNE

EESSÕNA

Käesolev juhend "Standardite ISO/IEC 10646 ja Unicode kasutusjuhend" on heaks kiidetud Eesti Standardikeskuse tehnilises komitees EVS/TK 4 "Infotehnoloogia". Juhendi koostas Indrek Hein Eesti Keele Instituudist.

Märkide universaalse kodeerimisviisi standard sai alguse 1988. aastal ning seda arendavad paralleelselt kaks organisatsiooni – ISO (JTC1, SC 2 *Kodeeritud märgistikud*) ja Unicode Consortium. Käesolev rakendusjuhend käsitleb seda standardit ühtsena ja kasutab tinglikult nimetust Unicode. Koostamisel on kasutatud ISO standardi versiooni ISO/IEC 10646:2003 ja Unicode standardi versiooni 4, mis mõlemad ilmusid 2003. aastal. Vajadusel juhitakse tekstis tähelepanu, kui üks või teine termin või määratlus on kasutusel ainult ühes neist standardeist.

ISO/IEC 10646 kannab ühtset pealkirja *Infotehnoloogia – universaalne mitme-oktetiline kodeeritud märgistik (UCS).* Lisad A kuni D on normatiivsed ISO/IEC 10646 osad, lisad E kuni S teatmelised.

ISO/IEC standardit haldab ISO/IEC JTC1/SC2, keda Eestis esindab Standardikeskus.

> Märkus. ISO/IEC 10646 tulevased väljaanded ei muuda praeguse versiooniga paika pandud märkide nimesid ja asukohti.

Juhend on kinnitatud ja kasutusele võetud Eesti Standardikeskuse 16.03.2005 käskkirjaga nr 24.

See dokument on EVS-i poolt loodud eelvaade

© Kirjastamis- ja paljundusõigus kuulub Eesti Standardikeskusele

SISUKORD

SISSEJUHATUS 1

1 KÄSITLUSALA 1

2 NORMATIIVVIIDE 1

3 TERMINID JA MÄÄRATLUSED 2

4 KOOSTAMISPÕHIMÕTTED 5

5 UCSI ÜLDSTRUKTUUR 6
5.1 Struktuur 6
5.1.1 Koodpositsioonide lühitähised (UID) 8
5.1.2 UCS järjenditähised 9
5.2 Koodpositsioonide sisu 9
5.3 Märkide paigutus 11
5.4 Märkide kodeerimine 12
5.4.1 UTF-32 13
5.4.2 UTF-16 13
5.4.3 UTF-8 14
5.4.3.1 UTF-8 koodi moodustamine 14
5.4.3.2 UTF-8 ja UTF-32 vahelised teisendused 16
5.5 Teksti kodeerimine 18

6 STANDARDILE VASTAVUS 19
6.1 Üldised vastavusnõuded 19

7 MÄRKIDE OMADUSED 20
7.1 Unicode'i märkide andmebaas 20
7.2 Tõste 21
7.3 Kombineerumisklassid 21
7.4 Üldkategooriad 21
7.5 Arvväärtused 22
7.6 Märkide peegeldatud kujud kahesuunalise kirja kontekstis 23
7.7 Erimärgid 23
7.8 Ühilduvusmärgid 23
7.9 Märkide nimetamine 23

8 ALAMHULGAD 24
8.1 Piiratud alamhulk 24
8.2 Valitud alamhulk 24

9 RAKENDUSTASEMED 24
9.1 1. rakendustase 24
9.2 2. rakendustase 24
9.3 3. rakendustase 24

10 JUHTFUNKTSIOONIDE KASUTAMINE UCS-IGA 25

See dokument on EVS-i poolt loodud eelvaade

11 IDENTIFITSEERIVA TEABE DEKLAREERIMINE 25
11.1 Identifitseeriva teabe eesmärk ja kontekst 25
11.2 UCS-i kodeeritud esituskuju ja rakendustaseme identifitseerimine 26

12 KOODITABELITE JA LOENDITE STRUKTUUR 26

13 KOMBINEERUVAD MÄRGID 27
13.1 Kombineeruvate märkide asukoht 27
13.2 Kuju kooditabelites 27
13.3 Kombineeruvate märkide kombinatsioonid 27
13.4 Märgikogud, mis sisaldavad kombineeruvaid märke 29
13.5 Variandivalijad (variation selectors) 29

Lisa A (normatiivlisa) Graafiliste märkide märgikogud alamhulkade moodustamiseks 30
A.1 Kodeeritud graafiliste märkide kogud 30
A.2 Muud BMP kogud 31
A.2.1 281 MES-1 31
A.2.2 282 MES-2 31

Lisa B (normatiivlisa) Kombineeruvate märkide nimekiri 33
B.1 Kombineeruvate märkide koondnimekiri 33

Lisa F (teatmelisa) Kujuvalikumärgid (Alternate format characters) 34

Lisa H (teatmelisa) "Signatuuride" kasutamine UCSi identifitseerimiseks 35

Lisa L (teatmelisa) Juhtnöörid märkide nimede moodustamiseks 36
Reegel 1 36
Reegel 2 36
Reegel 3 37
Reegel 4 37
Reegel 5 37
Reegel 6 37
Reegel 7 38
Reegel 8 38
Reegel 9 38
Reegel 10 39
Reegel 11 39
Reegel 12 39
Reegel 13 39

Lisa N (teatmelisa) Viitamine välistele tähevalimitele 40
N.1 Tähevalimitele ja nende kodeeringule viitamise meetodid. 40
N.2 ASN.1 abstraktse märgisüntaksi identifitseerimine 40
N.3 ASN.1 märgiedastussüntaksi identifitseerimine 41

Lisa P (teatmelisa) Lisateavet märkide kohta 43

See dokument on EVS-i poolt loodud eelvaade

EESTI STANDARD EVS JUHEND 8:2005

STANDARDITE ISO/IEC 10646 JA UNICODE KASUTUSJUHEND

Guidelines for using of ISO/IEC 10646 and UNICODE

See dokument on tasuta koolitusmoodulite eelvaade

SISSEJUHATUS

ISO/IEC 10646 (Unicode) määratleb universaalse mitmeoktetilise kodeeritud märgistiku (**UCS**), mis on kasutatav kõigi maailma kirjakeelte (kirjasüsteemide) ning täiendavate sümbolite esitamiseks, edastamiseks, andmevahetuseks, töötlemiseks, salvestuseks, sisestuseks ning kuvamiseks. Lisaks määratletakse UCS kodeerimisviisid, sortimise põhimõtted, märkide kanoonilised kujud jpm. Unicode kuulub juba lahutamatu osana Interneti juurde, olles HTML ja XML standardite kasutatavaks märgistikuks. Unicode'i toetavad operatsioonisüsteemid, andmebaasid, programmeerimiskeeled, s.t kokkuvõttes kogu IT infrastruktuur.

Unicode kui kõigi seniste märgistike ülemhulk on täitmas oma algset eesmärki – asendada kaheksabitiste kooditabelite kirev ja omavahel kokkusobimatu paljusus ühtse kodeeringuga, mis käsitleb standardselt kõiki kasutusel olevaid märke ja kirjaviise. Unicode võimaldab kodeerida üle miljoni märgi, mis on enam kui piisav nii kõigi kasutusel olevate kui surnud keelte tähestike ja märkide kodeerimiseks. Standardi praeguses versioonis on kodeeritud 96382 märki (üle 70000 neist hieroglüüfid), järgmised versioonid lisavad märke veelgi. Rakenduslik huvi piirdub enamasti Unicode'i esimese 65535 märgi ehk UCSi **mitmekeelse alustasandi** ehk **BMP**-ga.

> Märgile viitamiseks kasutatakse tema koodpunkti, kuueteistkümnendarvu vahemikus 0..10FFFF, lisades kokkuleppeliselt ette "U+" (täpsemalt vt 2.1.1).

1 KÄSITLUSALA

Juhend käsitleb ainult neid Unicode lisasid (säilitades numeratsiooni), mis Eesti kasutajat otsesemalt puudutavad. Vaatluse alt jäävad välja märkide nimetamise juhendid, paremalt vasakule kirjutamisega ja hieroglüüfidega seotud probleemistik jms, samuti mahupiirangu tõttu märgitabelid. Soovijad võivad nende osadega tutvuda veebileheküljel www.unicode.org.

2 NORMATIIVVIIDE

Käesolev juhend viitab järgmisele dokumendile:

ISO/IEC 10646 *Infotehnoloogia – universaalne mitmeoktetiline kodeeritud märgistik (UCS).* Lisad A kuni D on normatiivsed ISO/IEC 10646 osad.