

Een pagina aangesproken met codes van één byte

Elke letter op de drie regels hieronder kost één byte van de paginastroom van dit bestand in plaats van twee. Een lettertype wordt bereikt via glyphindex, tenzij anders aangegeven, twee bytes per glyph, waardoor elke glyph van de letter bereikbaar is. Via een codering wordt het in plaats daarvan aangesproken met codes van één byte: hoogstens 255, elk voor één teken, en het bestand zegt voor welk teken elke code staat.

Gezet met codes van één byte

HQF Development

Een factuur, een loonstrook, een pakbon.

Eén byte per letter, en het bestand zegt welke.

De codes die het vastlegde

De codering is afgestemd op de woorden die deze pagina tekent: de spatie houdt code 32, en elk ander teken krijgt de laagste vrije code vanaf 33, dus niets in de stroom is een byte die leessoftware moet escaperen. De eerste zijn H 33 · Q 34 · F 35 · D 36 · e 37 · v 38 · l 39 · o 40.

Wat de stroom weegt

Getekend via glyphindex kosten de drie regels hierboven 546 bytes paginastroom; via codes van één byte 342 bytes, dat is 204 minder. Beide getallen zijn gemeten door het programma dat deze pagina tekende, niet met de hand geschreven.

Wat de losse byte 32 oplevert

Woorden die leessoftware uit elkaar kan duwen.

Woorden die leessoftware uit elkaar kan duwen.

De twee regels hierboven bevatten dezelfde tekens. De tweede is getekend met woordafstand ingesteld, die de pen verplaatst bij de losse byte 32 en bij geen andere code. Onder glyphindices is geen enkele code een losse byte 32, dus heeft de operator niets te bereiken en bewegen de woorden nooit; de spatie van deze codering staat op 32, dus bewegen ze wel.

Een codetabel genoemd bij haar norm

Счёт-фактура № 117 — оплата до 30 сентября

De regel hierboven is gezet via Windows-codetabel 1251, één keer genoemd in plaats van code voor code uitgeschreven: het bestand noemt de Windows-codering en noemt daarbovenop elke code die de tabel anders invult, dus een lezer vindt elke Cyrillische letter waar de tabel haar zet. De eerste bytes zijn D1 F7 B8 F2 2D F4 E0 EA.

Eén byte bereikt hoogstens 255 tekens, dus een pagina met meerdere schriften wordt nog steeds via glyphindex bereikt. Ook een ligatuur of kleinkapitaal wordt niet via een codering getekend: dat zijn glyphs waar geen teken voor staat, en een code staat voor een teken. Een bestand dat archiefconformiteit claimt, noemt de Windows-codering als die waartegen zijn codes gelezen worden, en dit bestand doet dat.

Dertig coderingen waartegen een code gelezen kan worden

Waar de byte 0xC9 in elk voor staat: het codepunt van het teken, en het teken zelf waar deze letter het kan tekenen.

WinAnsi	U+00C9 É	ISO 8859-8	niets	Windows-1251	U+0419 Ў
Standard	niets	ISO 8859-9	U+00C9 É	Windows-1253	U+0399 І
PDFDoc	U+00C9 É	ISO 8859-10	U+00C9 É	Windows-1254	U+00C9 É
ISO 8859-1	U+00C9 É	ISO 8859-11	U+0E29	Windows-1255	U+05B9 `
ISO 8859-2	U+00C9 É	ISO 8859-13	U+00C9 É	Windows-1256	U+0629 ̂
ISO 8859-3	U+00C9 É	ISO 8859-14	U+00C9 É	Windows-1257	U+00C9 É
ISO 8859-4	U+00C9 É	ISO 8859-15	U+00C9 É	Windows-1258	U+00C9 É
ISO 8859-5	U+0429 Ш	ISO 8859-16	U+00C9 É	Mac Roman	U+2026 ...
ISO 8859-6	U+0629 ̂	Windows-874	U+0E29	IBM 037	U+0049 І
ISO 8859-7	U+0399 І	Windows-1250	U+00C9 É	IBM 500	U+0049 І

Een codering met de hand vastgelegd

code 0x41 staat voor U+0416, waarvan de glyph Zhecyrillic heet

code 0x42 staat voor U+03A9, waarvan de glyph Omegagreek heet

code 0x43 staat voor U+20AC, waarvan de glyph Euro heet

code 0x44 staat voor U+00C9, waarvan de glyph Eacute heet

ЖΩ€É

De regel hierboven is getekend met alleen die vier codes. De codering vertrekt van helemaal geen codering, en het bestand noemt 5 codes die ervan afwijken.

Een letter zonder de ligatuur U+FB01 krijgt „fi” in de plaats, en een zonder het ohmteken U+2126, „Ω”.

Wat een letter tekent voor een teken dat hij niet kan tekenen, is zijn glyph 0, die elke subset behoudt.