

# Eine Seite, über Ein-Byte-Codes angesprochen

Jeder Buchstabe in den drei Zeilen unten kostet ein Byte im Seitenstrom dieser Datei statt zwei. Eine Schrift wird über den Glyphenindex angesprochen, sofern ihr nichts anderes gesagt wird, zwei Bytes je Glyph, womit jeder Glyph des Schnitts erreichbar ist. Über eine Kodierung erreicht, wird sie stattdessen mit Ein-Byte-Codes angesprochen: höchstens 255, jeder steht für ein Zeichen, und die Datei gibt an, für welches.

Über Ein-Byte-Codes gesetzt

HQF Development

Eine Rechnung, eine Gehaltsabrechnung, ein Lieferschein.

Ein Byte je Buchstabe, und die Datei sagt, welcher.

Die festgelegten Codes

Die Kodierung ist auf die Wörter zugeschnitten, die diese Seite zeichnet: Das Leerzeichen behält Code 32, und jedes andere Zeichen nimmt den niedrigsten freien Code ab 33, also steht im Strom kein Byte, das eine Lesesoftware maskieren müsste. Die ersten davon sind H 33 · Q 34 · F 35 · D 36 · e 37 · v 38 · l 39 · o 40.

Was der Strom wiegt

Über den Glyphenindex gezeichnet, kosten die drei Zeilen oben 626 Bytes Seitenstrom; über Ein-Byte-Codes 382 Bytes, also 244 weniger. Beide Zahlen misst das Programm, das diese Seite erstellt hat, keine ist von Hand eingetragen.

Was das einzelne Byte 32 bringt

Wörter, die eine Lesesoftware auseinanderchieben kann.

Wörter, die eine Lesesoftware auseinanderchieben kann.

Die beiden Zeilen oben enthalten dieselben Zeichen. Die zweite ist mit Wortabstand gezeichnet, der den Stift beim einzelnen Byte 32 weiterschiebt und bei keinem anderen Code. Unter Glyphenindizes ist kein Code ein einzelnes Byte 32, also findet der Operator nichts und die Wörter bewegen sich nie; das Leerzeichen dieser Kodierung liegt auf 32, also bewegen sie sich.

Eine Codepage, nach ihrer Norm benannt

Счёт-фактура № 117 — оплата до 30 сентября

Die Zeile oben ist über die Windows-Codepage 1251 gesetzt, einmal benannt, statt Code für Code ausgeschrieben: Die Datei nennt die Windows-Kodierung und darüber jeden Code, den die Codepage anders belegt, also findet ein Leser jeden kyrillischen Buchstaben dort, wo die Codepage ihn hinsetzt. Ihre ersten Bytes sind D1 F7 B8 F2 2D F4 E0 EA.

Ein Byte reicht für höchstens 255 Zeichen zugleich, also wird eine Seite mit mehreren Alphabeten weiter über den Glyphenindex erreicht. Auch Ligaturen und Kapitälchen lassen sich nicht über eine Kodierung zeichnen: Das sind Glyphen, für die kein Zeichen steht, und ein Code steht für ein Zeichen. Eine Datei, die Archivkonformität beansprucht, nennt die Windows-Kodierung als die, nach der ihre Codes gelesen werden, wie diese es tut.

## Dreißig Kodierungen, nach denen ein Code gelesen werden kann

Wofür das Byte 0xC9 in jeder steht: der Codepunkt des Zeichens und das Zeichen selbst, wo dieser Schnitt es zeichnen kann.

|            |          |              |          |              |            |
|------------|----------|--------------|----------|--------------|------------|
| WinAnsi    | U+00C9 É | ISO 8859-8   | nichts   | Windows-1251 | U+0419 Ё   |
| Standard   | nichts   | ISO 8859-9   | U+00C9 É | Windows-1253 | U+0399 Ĳ   |
| 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 ˆ   |
| 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 Щ | ISO 8859-16  | U+00C9 É | Mac Roman    | U+2026 ... |
| ISO 8859-6 | U+0629 ̐ | Windows-874  | U+0E29   | IBM 037      | U+0049 Ĳ   |
| ISO 8859-7 | U+0399 Ĳ | Windows-1250 | U+00C9 É | IBM 500      | U+0049 Ĳ   |

### Eine von Hand festgelegte Kodierung

Code 0x41 steht für U+0416, dessen Glyph Zhecyrillic heißt

Code 0x42 steht für U+03A9, dessen Glyph Omegagreek heißt

Code 0x43 steht für U+20AC, dessen Glyph Euro heißt

Code 0x44 steht für U+00C9, dessen Glyph Eacute heißt

ЖΩ€É

Die Zeile oben ist allein über diese vier Codes gezeichnet. Die Kodierung geht von überhaupt keine Kodierung aus, und die Datei nennt 5 Codes, die davon abweichen.

Ein Schnitt ohne die Ligatur U+FB01 bekommt an ihrer Stelle „fi“, und einer ohne das Ohm-Zeichen U+2126 „Ω“.

Was ein Schnitt für ein Zeichen zeichnet, das er nicht zeichnen kann, ist sein Glyph 0, den jede Teilmenge behält.