Une page adressée par des codes d’un octet

Chaque lettre des trois lignes ci-dessous coûte un octet du flux de cette page, et non deux. Une police est atteinte par indice de glyphe tant qu’on ne précise rien d’autre, à deux octets par glyphe, ce qui rend accessibles tous les glyphes qu’elle contient. Atteinte par un encodage, elle est adressée par des codes d’un octet : 255 au plus, chacun représentant un caractère, et le fichier indique quel caractère chacun d’eux représente.

Composé avec des codes d’un octet

HQF Development

Une facture, une fiche de paie, un bon de livraison.

Un octet par lettre, et le fichier dit laquelle.

Les codes retenus

L’encodage est ajusté aux mots de cette page : l’espace garde le code 32, et chaque autre caractère prend le plus petit code libre à partir de 33, si bien qu’aucun octet du flux ne demande d’échappement au logiciel de lecture. Les premiers sont H 33 · Q 34 · F 35 · D 36 · e 37 · v 38 · l 39 · o 40.

Ce que pèse le flux

Écrites par indice de glyphe, les trois lignes ci-dessus coûtent 598 octets de flux de page ; écrites par des codes d’un octet, 368 octets, soit 230 de moins. Les deux chiffres sont mesurés par le programme qui a créé cette page, et non saisis à la main.

Ce que permet l’octet 32

Des mots qu’un logiciel de lecture peut écarter.

Des mots qu’un logiciel de lecture peut écarter.

Les deux lignes ci-dessus contiennent les mêmes caractères. La seconde est écrite avec un espacement des mots, qui avance la position d’écriture sur l’octet 32 et sur aucun autre code. Par indice de glyphe, aucun code ne vaut l’octet 32 seul : l’opérateur n’a rien à atteindre et les mots ne bougent jamais ; l’espace de cet encodage est en 32, donc ils bougent.

Une page de code nommée par sa norme

Счёт-фактура № 117 — оплата до 30 сентября

La ligne ci-dessus est composée avec la page de code Windows 1251, nommée une fois au lieu d’être détaillée code par code : le fichier annonce l’encodage Windows et redéfinit par-dessus chaque code que la page de code remplit autrement, si bien qu’un logiciel de lecture trouve chaque lettre cyrillique là où elle la place. Ses premiers octets sont D1 F7 B8 F2 2D F4 E0 EA.

Un octet adresse 255 caractères à la fois : une page qui mêle plusieurs alphabets reste atteinte par indice de glyphe. Aucune ligature ni petite capitale ne passe non plus par un encodage : ce sont des glyphes qu’aucun caractère ne représente, et un code représente un caractère. Un fichier qui se déclare conforme à l’archivage annonce l’encodage Windows comme référence de lecture de ses codes, ce que fait celui-ci.

Trente et un encodages selon lesquels lire un code

Ce que l’octet 0xC9 représente dans chacun : le point de code du caractère, et le caractère lui-même là où cette police sait le dessiner.

WinAnsi	U+00C9 É	ISO 8859-9	U+00C9 É	Windows-1254	U+00C9 É
Standard	rien	ISO 8859-10	U+00C9 É	Windows-1255	U+05B9 ֹ
PDFDoc	U+00C9 É	ISO 8859-11	U+0E29	Windows-1256	U+0629 ة
ISO 8859-1	U+00C9 É	ISO 8859-13	U+00C9 É	Windows-1257	U+00C9 É
ISO 8859-2	U+00C9 É	ISO 8859-14	U+00C9 É	Windows-1258	U+00C9 É
ISO 8859-3	U+00C9 É	ISO 8859-15	U+00C9 É	Mac Roman	U+2026 …
ISO 8859-4	U+00C9 É	ISO 8859-16	U+00C9 É	IBM 037	U+0049 I
ISO 8859-5	U+0429 Щ	Windows-874	U+0E29	IBM 500	U+0049 I
ISO 8859-6	U+0629 ة	Windows-1250	U+00C9 É	Mac Expert	U+F7B8
ISO 8859-7	U+0399 Ι	Windows-1251	U+0419 Й		
ISO 8859-8	rien	Windows-1253	U+0399 Ι		

Un encodage réglé à la main

le code 0x41 représente U+0416, dont le glyphe s’appelle Zhecyrillic
le code 0x42 représente U+03A9, dont le glyphe s’appelle Omegagreek
le code 0x43 représente U+20AC, dont le glyphe s’appelle Euro
le code 0x44 représente U+00C9, dont le glyphe s’appelle Eacute

ЖΩ€É

La ligne ci-dessus n’est tracée qu’avec ces quatre codes. Base de l’encodage : aucun encodage ; le fichier nomme 5 codes qui s’en écartent.

Une police sans la ligature U+FB01 reçoit « fi » à sa place, et une police sans le signe ohm U+2126, « Ω ».

Pour un caractère qu’elle ne sait pas dessiner, une police dessine son glyphe 0, que tout sous-ensemble conserve.