Unicode explained

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 27 octobre 2007

https://www.bortzmeyer.org/unicode-explained.html

Auteur(s) : Jukka Korpela
ISBN n°0-596-10121-x
Éditeur : O'Reilly
Publié en 2006

Voici le premier livre chez O'Reilly à propos d'Unicode. C'est un gros pavé de 700 pages et il ne couvre pourtant pas tout.

Le monde des écritures humaines est très riche et très compliqué et donc, logiquement, Unicode est compliqué. La norme est bien écrite, certes, mais rien ne peut rendre simple une réalité aussi complexe. N'attendons donc pas de ce livre des miracles, il explique déjà très bien beaucoup de choses.

L'orientation générale est plutôt vers l'auteur de documents. Le programmeur sera frustré, seul un chapitre lui est consacré, et une bonne partie est occupée par des langages de balisage comme HTML, qui ne sont pas de la programmation.

En revanche, Jukka Korpela <http://www.cs.tut.fi/~jkorpela/personal.html> explique avec beaucoup de détails (chapitres 4 à 6) la structure de la norme Unicode, les différents termes utilisés par celle-ci, les encodages et les innombrables informations qu'on peut trouver dans la base de données Unicode <https://www.bortzmeyer.org/unicode-to-sql.html>. Plus original, on trouve un excellent chapitre 2 consacré aux techniques permettant de saisir des caractères Unicode. Bien qu'il soit nettement trop orienté Windows à mon goût, ce chapitre réussit très bien à décrire les différentes méthodes.

Pour renforcer le côté « livre conçu pour les auteurs de textes, pas pour les programmeurs », je signale l'excellent chapitre 8, sur les usages des caractères, où on apprend plein de choses, même sur l'humble signe +. Beaucoup de caractères ont en effet plusieurs usages (la lettre grecque oméga - U+03A9 ou [Caractère Unicode non montré [1]] - sert également à noter l'ohm, l'astérisque des énumérations en anglais sert également aux programmeurs à noter la multiplication, etc).

1. Car trop difficile à faire afficher par LaTeX

Le chapitre 11, consacré à la programmation est, on l'a dit, un peu limité, et il vaut sans doute mieux (re)lire "*Unicode demystified*" <https://www.bortzmeyer.org/unicode-demystified.html>, de Richard Gillam.

Le chapitre 10 (sur les protocoles Internet) couvre également le problème intéressant d'écrire des textes en Unicode pour les publier sur Internet. L'auteur fait justement remarquer que, si on écrit en Unicode (par exemple en UTF-8), on sera tenté d'utiliser beaucoup plus de caractères, peut-être au détriment de la lisibilité du texte par les destinataires. Si on a donc relativement peu de caractères non-latins, il suggère de taper le texte en ASCII ou en ISO 8859 et d'utiliser les échappements comme `&#x0627;` (alif, [Caractère Unicode non montré]). Curieusement, c'est justement ce que fait aujourd'hui l'auteur de ce blog, qui, en raison de problèmes avec UTF-8 <https://www.bortzmeyer.org/pas-encore-utf8.html> rédige en Latin-1 et met les caractères non-latins avec des entités numériques XML.

Le style est très bon et les explications claires. La qualité générale de ce livre est celle des autres publications de cet éditeur et ce livre est donc à recommander aux non-programmeurs qui veulent comprendre les détails d'Unicode (ceux qui parlent français préféreront peut-être le livre d'Andries <https://www.bortzmeyer.org/unicode-en-pratique.html>).