

Unicode 5.0 en pratique

Stéphane Bortzmeyer

<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 17 mai 2008

<http://www.bortzmeyer.org/unicode-en-pratique.html>

Auteur(s) : Patrick Andries

ISBN n°978-2-10-051140-2

Éditeur : Dunod

Publié en 2008

Unicode est un jeu de caractères permettant de représenter toutes les écritures du monde. Il serait donc logique qu'il soit documenté dans plusieurs langues mais, sauf erreur, ce livre de Patrick Andries est le premier à parler d'Unicode en français.

Le sujet est très riche et très complexe. Certaines personnes, ignorant l'ampleur du problème, disent parfois qu'Unicode est « trop complexe ». La réalité est que le monde est complexe; Unicode choisissant de suivre les usages existants (au lieu que les usages se plient aux limites de la technique), il doit traiter d'innombrables cas et le résultat est certainement complexe, mais pas inutilement complexe. Patrick Andries connaît bien Unicode pour être membre du Consortium Unicode depuis des années et pour être un promoteur de cette norme en français (son site <<http://hapax.qc.ca/>> regorge d'informations sur Unicode). Il s'est donc attaqué à la tâche d'écrire le premier livre en français sur Unicode.

Cela suppose quelques défis, notamment de trouver des traductions correctes et compréhensibles pour certains termes. Je n'ai pas reconnu du premier coup le "*word joiner*" (U+2060) derrière le « gluon de mots » et les « seizets d'indirection » m'ont également dérouté, même si le terme est bien plus correct que l'anglais "*surrogates*". D'une manière générale, ce livre est peut-être un peu trop difficile à lire à force de vouloir être exact et de toujours utiliser des termes corrects, en évitant raccourcis et approximations. Il paraît que la langue française est bien adaptée à cela...

Le plan du livre est classique; concepts de base sur les écritures (chapitre 1); les jeux de caractères (chapitre 2); les concepts de base d'Unicode (chapitres 3 et 4); grand tour d'horizon des caractères dont on dispose dans Unicode (chapitres 5 à 8); applications d'Unicode, notamment dans le monde des protocoles et formats Internet (chapitres 9 à 11); l'internationalisation de logiciels (chapitre 12, très intéressant mais dont le lien avec Unicode est ténu); Unicode et les polices (chapitre 13, dont je n'ai vu aucun équivalent dans les autres livres sur Unicode).

Les exemples de programmation du chapitre 12 sont presque tous empruntés à C# ou à Java, des langages plutôt « costard / cravate », qui ont souvent l'avantage d'un grand nombre de bibliothèques pour manipuler l'Unicode (par exemple via un accès facile à ICU) et le défaut d'un modèle de caractères peu Unicodien (par exemple, dans Java, les caractères en dehors du Plan Multilingue de Base ne sont pas des caractères Unicode comme les autres, contrairement à Python ou Haskell). Pour avoir une petite idée du support Unicode dans d'autres langages, on peut regarder mon article Unicode en divers langages <<http://www.bortzmeyer.org/unicode-en-divers-langages.html>>.

De même, la plupart des exemples concrets, par exemple sur la saisie de caractères Unicode, ne concernent que MS-Windows et les utilisateurs d'Unix qui ont des difficultés avec Unicode <<http://www.bortzmeyer.org/pas-encore-utf8.html>> ne trouveront pas beaucoup à se mettre sous la dent.

L'écriture même d'un tel livre est un exploit, dans un monde où les systèmes informatiques sont encore rares d'avoir un parfait support d'Unicode. Peu de logiciels permettent de composer un ouvrage avec d'innombrables exemples Unicode (LaTeX en est hélas peu capable <<http://www.bortzmeyer.org/latex-unicode.html>>). Le livre a été écrit avec MS-Word, puis porté sur InDesign pour l'édition finale. InDesign s'est parfois montré trop intelligent, par exemple en remplaçant des caractères qu'il pensait inadaptés par d'autres. C'est ainsi qu'un exemple discutant le guillemet anglo-saxon, " (U+0022), présente en fait un guillemet français, le « (U+00AB). Une page Web d'errata <<http://hapax.qc.ca/errata-unicode-5.html>> présente les erreurs recensées dans le livre.

Au final, c'est un livre que je recommande chaudement. Les programmeurs qui veulent mettre en œuvre Unicode (et pas seulement l'utiliser) trouveront plus d'informations dans le livre de Gillam <<http://www.bortzmeyer.org/unicode-demystified.html>> mais ceux qui veulent se servir d'Unicode dans un contexte Internet/XML seront au contraire bien plus satisfaits avec le livre d'Andries (qui est plus proche, par le public visé, du livre - en anglais - de Korpela <<http://www.bortzmeyer.org/unicode-explained.html>>).