Il n'existe pas de « caractères spéciaux »

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 6 janvier 2011

http://www.bortzmeyer.org/pas-de-caracteres-speciaux.html

Ceci est un billet d'humeur. Si vous n'aimez pas les billets d'humeur, ou bien si vous trouvez qu'ils sont un moyen facile de remplir son blog à peu de frais, passez à l'entrée suivante dans votre flux de syndication. La raison de ce billet est qu'une recherche sur Google des termes « caractères spéciaux <http://www.google.com/search?ie=UTF-8&q=%22caract%C3%A8res+sp%C3%A9ciaux%22> » donne pour moi 201 000 résultats, presque tous utilisant cette expression à tort. Or, **il n'y a pas de caractères spéciaux. Tous les caractères Unicode sont égaux !**

Un exemple parmi ces 201 000, un article d'une revue de bonne qualité, consacré à une bogue informatique amusante <http://www.60millions-mag.com/actualites/archives/quand_l_iphone_se_met_a_begayer> et qui explique « l[Caractère Unicode non montré [1]]iPhone aurait du mal à gérer les caractères spéciaux, type [Caractère Unicode non montré]ç[Caractère Unicode non montré] ou [Caractère Unicode non montré]ê[Caractère Unicode non montré] insérés dans les SMS ». Mais ces caractères n'ont rien de spécial ! Ce sont des caractères Unicode comme les autres, et, dans un article en français, le ç n'a rien de plus spécial que le c ! Les ignorants emploient en général ce terme de « caractères spéciaux » pour parler de ceux qui ne sont pas dans US-ASCII. Mais c'est un provincialisme très étroit : même pour les langues utilisant l'alphabet latin, la quasi-totalité d'entre-elles utilisent des caractères qui ne sont **pas** dans US-ASCII et ne sont pas spéciaux pour autant.

En Europe, je crois que seuls l'anglais, le frison et le néerlandais peuvent s'écrire avec uniquement ASCII. Et encore : en anglais, des mots comme "*résumé*" <http://en.wiktionary.org/wiki/r%C3%A9sum%C3%A9#English> (à ne pas confondre avec le verbe "*resume*" <http://en.wiktionary.org/wiki/resume#English>, qui n'a pas le même sens, c'est pour cela que les accents sont importants) nécessitent aussi ces caractères injustement qualifiés de spéciaux. De même, en néerlandais, on a des noms propres avec accent (comme "*België*"), certains utilisent <http://meinamsterdam.nl/nouveau-mot-ij> la ligature [Caractère Unicode non montré] (U+0133) ou bien on trouve un accent dans le numéral 1 <http://en.wiktionary.org/wiki/%C3%A9%C3%A9n#Dutch>. (Autres exemples sur Wikipédia.)

1. Car trop difficile à faire afficher par LaTeX

Donc, je le répète, tous les caractères Unicode sont égaux : ils ont tous un **point de code** (un nombre, comme U+00E7 pour ç, U+0063 pour c, U+00E9 pour é ou U+0178 pour [Caractère Unicode non montré]) et un nom. (Seule exception, dans le monde Unicode, l'encodage UCS-2, qui ne traite pas les points de code supérieux à 65535, pourtant la majorité des caractères Unicode. Il ne devrait donc jamais être employé. UTF-16, lui, traite d'une manière... spéciale, ces caractères mais, au moins, il permet de les représenter. Mauvais encodage, toutefois.)

Les seuls caractères qu'on peut légitimement appeler « caractères spéciaux » sont ceux qui nécessitent un échappement dans le langage qu'on utilise. Ainsi, il est normal de dire « \ est un caractère spécial dans les chaînes de caractères en C » ou bien « & est un caractère spécial en XML » et de rappeler aux programmeurs qu'ils doivent faire attention en les manipulant <http://www.bortzmeyer.org/creer-xml-par-programme.html>. Mais c'est tout.

Une note au passage sur les SMS, qui étaient cités dans l'exemple que je donnais : la question de l'Unicode dans SMS est complexe et, dans ce cas précis, certains caractères sont en effet spéciaux et traités différemment. Mais la très grande majorité des 201 000 autres exemples ne concernent pas cette technologie archaïque.

Merci à Alix Guillard, Marco Davids, Patrick Vande Walle, André Sintzoff et Rafaël Garcia-Suarez pour leurs commentaires érudits sur le néerlandais.