

Unicode 5.1

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 7 Avril 2008

<http://www.bortzmeyer.org/unicode-5.1.html>

Le consortium Unicode vient d'annoncer la sortie de la version 5.1 de la célèbre norme de jeu de caractères. 1 624 caractères de plus, permettant à Unicode de dépasser les 100 000 caractères.

Parmi les nouveaux venus, un certain nombre de caractères arabes (notamment pour écrire le persan) ou cyrilliques, le Eszett allemand en majuscule (la précédente majuscule de [Caractère Unicode non montré ¹] était la chaîne SS, ce sera désormais le caractère Unicode U+1E9E, que votre navigateur ne sait probablement pas encore afficher), l'alphabet Sundanese (que je ne connaissais pas), d'autres alphabets nouveaux comme le Cham, les caractères du Mah-jong (cf. le roman « Sous les vents de Neptune » de Fred Vargas), et plein d'autres caractères rigolos comme U+2B3D, « LEFTWARDS TWO-HEADED ARROW WITH TAIL WITH DOUBLE VERTICAL STROKE » ou U+2E2A, « TWO DOTS OVER ONE DOT PUNCTUATION ».

La majuscule du Eszett a suscité bien des controverses (cf. <http://de.wikipedia.org/wiki/Versal-Eszett>), certains germanophones expliquant que cette lettre n'existe pas. On peut consulter la proposition initiale en <http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3227.pdf>, avec des exemples d'utilisation, pris par exemple sur des pierres tombales.

Le programme de conversion de la base Unicode en SQL (<http://www.bortzmeyer.org/unicode-to-sql.html>) est un excellent outil pour explorer cette nouvelle version :

```
ucd=> SELECT count(*) AS Total FROM Characters;                                total
-----
100713
(1 row)

ucd=> SELECT To_U(codepoint) AS Codepoint, name FROM
      Characters WHERE version = '5.1';
codepoint | name
-----+-----
```

1. Car trop difficile à faire afficher par L^AT_EX

U+1E9E | LATIN CAPITAL LETTER SHARP S
U+A66E | CYRILLIC LETTER MULTIOCULAR O
...
U+372 | GREEK CAPITAL LETTER ARCHAIC SAMPI
U+373 | GREEK SMALL LETTER ARCHAIC SAMPI
...
U+520 | CYRILLIC CAPITAL LETTER EL WITH MIDDLE HOOK
U+521 | CYRILLIC SMALL LETTER EL WITH MIDDLE HOOK
U+522 | CYRILLIC CAPITAL LETTER EN WITH MIDDLE HOOK
U+523 | CYRILLIC SMALL LETTER EN WITH MIDDLE HOOK
U+606 | ARABIC-INDIC CUBE ROOT
U+607 | ARABIC-INDIC FOURTH ROOT
U+608 | ARABIC RAY
...