

Version 7 d'Unicode

Stéphane Bortzmeyer

<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 19 juin 2014

<https://www.bortzmeyer.org/unicode-7-0.html>

Le 16 juin a vu la sortie d'une nouvelle version <<http://unicode-inc.blogspot.fr/2014/06/announcing-unicode-standard-version-70.html>> du jeu de caractères Unicode, la 7.0. On peut trouver une description des principaux changements en <<http://www.unicode.org/versions/Unicode7.0.0/>> mais voici ceux qui m'ont intéressé particulièrement. (Il n'y a pas de changement radical.)

Pour explorer plus facilement la grande base Unicode, j'utilise un programme qui la convertit en SQL <<https://www.bortzmeyer.org/unicode-to-sql.html>> et permet ensuite de faire des analyses variées. Faisons quelques requêtes SQL :

```
ucd=> SELECT count(*) AS Total FROM Characters;
total
-----
113021
```

Plus de 113 000 caractères. Lesquels ont été apportés par la version 7?

```
ucd=> SELECT version,count(version) FROM Characters GROUP BY version ORDER BY version;
version | count
-----+-----
...
6.0    | 2088
6.1    |  732
6.2    |    1
6.3    |    5
7.0    | 2834
```

2 834 nouveaux. Lesquels?

```

ucd=> SELECT To_U(codepoint) AS Codepoint, name FROM Characters WHERE version='7.0';
codepoint | name
-----+-----
...
U+2B5E    | BENT ARROW POINTING DOWNWARDS THEN NORTH EAST
...
U+10350   | OLD PERMIC LETTER AN
...
U+10500   | ELBASAN LETTER A
...
U+10600   | LINEAR A SIGN AB001
...
U+10880   | NABATAEAN LETTER FINAL ALEPH
...
U+1236F   | CUNEIFORM SIGN KAP ELAMITE
...
U+1F57D   | RIGHT HAND TELEPHONE RECEIVER
...
U+1F6F2   | DIESEL LOCOMOTIVE
...

```

Comme on le voit, c'est varié. On trouve vingt-trois nouveaux alphabets, ayant souvent cessés d'être utilisés, et intégrés à Unicode surtout pour les recherches historiques, comme par exemple le vieux permien, l'elbasan, le nabatéen ou le Linéaire A, toujours pas déchiffré (ce qui explique les noms des caractères, qui sont des chiffres sans sémantique). Il y a aussi l'ajout de plus de mille nouveaux caractères cunéiformes.

La partie la plus spectaculaire, mais pas forcément la plus utile, est l'ajout de nombreux symboles, comme les innombrables flèches (comme la U+2B5E ci-dessus), ou la locomotive Diesel U+1F6F2... À ceux qui disent « digital » pour numérique (comme les innombrables « agences digitales » qui font de la « stratégie digitale »), on pourra désormais répondre avec l'image du caractère U+1F595, « *REVERSED HAND WITH MIDDLE FINGER EXTENDED* »... Pour mieux apprécier ces nouveaux symboles (avec des représentations graphiques), vous pouvez regarder la liste de ceux de la série U+1F300 <<http://www.unicode.org/charts/PDF/U1F300.pdf>> et suivants... Si vous avez les bonnes polices de caractères, voici les caractères pris en exemple plus haut : [Caractère Unicode non montré ¹] [Caractère Unicode non montré] [Caractère Unicode non montré] [Caractère Unicode non montré] [Caractère Unicode non montré] [Caractère Unicode non montré] [Caractère Unicode non montré] [Caractère Unicode non montré] [Caractère Unicode non montré] [Caractère Unicode non montré]... Une autre solution est de se servir d'UniView <<http://people.w3.org/rishida/uniview/>> qui a été mis à jour avec les données de la version 7. Quant aux noms officiels en français de tous ces caractères, ils sont en ligne <<http://hapax.qc.ca/ListeDesNoms-7.0%282014-06-11%29.txt>>.

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