

Version 11 d'Unicode

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 6 juin 2018

<https://www.bortzmeyer.org/unicode-11-0.html>

Aujourd'hui 6 juin, la nouvelle version d'Unicode <<http://blog.unicode.org/2018/06/announcing-unicode.html>> est sortie, la 11.0. Une description officielle des principaux changements est disponible <<https://www.unicode.org/versions/Unicode11.0.0/>> 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
-----
137439
```

Combien de caractères sont arrivés avec la version 11?

```
ucd=> SELECT version,count(version) FROM Characters GROUP BY version ORDER BY version::float;
...
9.0      | 7500
10.0     | 8518
11.0     | 684
```

684 nouveaux, bien moins que dans les versions précédentes. Quels sont ces nouveaux caractères?

```

ucd=> SELECT To_U(codepoint) AS Codepoint, name FROM Characters WHERE version='11.0';
codepoint | name
-----+-----
...
U+1F9B8 | SUPERHERO
U+1F9B9 | SUPERVILLAIN
U+1F9C1 | CUPCAKE
...
U+10D00 | HANIFI ROHINGYA LETTER A
...
U+16E60 | MEDEFAIDRIN SMALL LETTER M
...
U+1D2E0 | MAYAN NUMERAL ZERO
...
U+1F12F | COPYLEFT SYMBOL
...
U+1F99D | RACCOON
U+1F99E | LOBSTER
U+1F99F | MOSQUITO
...
U+1F9B0 | EMOJI COMPONENT RED HAIR

```

Outre les habituels emojis plus ou moins utiles, et le symbole du copyleft (enfin!) qui plaira aux libristes, on trouve aussi six écritures plus ou moins nouvelles comme le medefaidrin, les chiffres mayas ou comme le hanifi. Les Rohingyas se font massacrer mais au moins leur écriture est désormais dans Unicode.

Toujours dans les emojis, on notera que la norme a précisé que les emojis n'ont pas forcément de genre <<http://www.unicode.org/reports/tr51/tr51-14.html#gender-neutral>>. Et elle a ajouté des modificateurs permettant de faire varier l'image comme le U+1F9B0 pour mettre des cheveux roux à un personnage, ou comme les changements de direction <<http://www.unicode.org/reports/tr51/tr51-14.html#Direction>>. Une des erreurs les plus souvent commises à propos des emojis (et d'ailleurs à propos d'Unicode en général) est de croire que l'image proposée par Unicode est normative : ce n'est qu'un exemple, et chaque auteur de police peut l'adapter (comme l'a récemment montré l'affaire de la salade Google <<https://www.numerama.com/pop-culture/383385-android-retire-loeuf-de-son-emoji-salade-pour-quelle-soit-vegan-et-plus-inclusif.html>>). Ainsi, si l'image proposée d'un coureur est un homme aux cheveux sombres, rien n'empêche une police Unicode d'utiliser une femme aux cheveux blonds. Pour les cas où il faut préciser, Unicode offre des mécanismes de modification d'un emoji comme les séquences ZWJ <http://www.unicode.org/reports/tr51/tr51-14.html#Emoji_ZWJ_Sequences>. Si elles sont gérées par votre logiciel (cela semble rare aujourd'hui dans le monde Unix libre mais ça marche, par exemple, chez Apple), vous devriez voir ici un coureur et une coureuse : [Caractère Unicode non montré ¹][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré]. Si vous voyez au contraire un personnage puis le symbole mâle ou femelle, c'est que votre logiciel ne traite pas ces séquences ZWJ. Voici ce que cela donne avec un Safari sur Mac :

Tiens, d'ailleurs, combien de caractères Unicode sont des symboles (il n'y a pas que les emojis parmi eux, mais Unicode n'a pas de catégorie « emoji ») :

```

ucd=> SELECT count(*) FROM Characters WHERE category IN ('Sm', 'Sc', 'Sk', 'So');
count
-----
7110

```

1. Car trop difficile à faire afficher par \LaTeX

Ou, en plus détaillé, et avec les noms longs des catégories :

```

ucd=> SELECT description,count(category) FROM Characters,Categories WHERE Categories.name = Characters.category
      description | count
-----+-----
Other_Symbol    |  5984
Math_Symbol     |   948
Modifier_Symbol |   121
Currency_Symbol |    57
(4 rows)

```

Si vous avez les bonnes polices de caractères, voici les caractères pris en exemple plus haut : [Caractère Unicode non montré]<<https://r12a.github.io/uniview/?char=1F9B8>>, [Caractère Unicode non montré]<<https://r12a.github.io/uniview/?char=1F9B9>>, [Caractère Unicode non montré]<<https://r12a.github.io/uniview/?char=1F9C1>>, [Caractère Unicode non montré]<<https://r12a.github.io/uniview/?char=10D00>>, [Caractère Unicode non montré]<<https://r12a.github.io/uniview/?char=16E60>>, [Caractère Unicode non montré]<<https://r12a.github.io/uniview/?char=1D2E0>>, [Caractère Unicode non montré]<<https://r12a.github.io/uniview/?char=1F12F>>, [Caractère Unicode non montré]<<https://r12a.github.io/uniview/?char=1F99D>>, [Caractère Unicode non montré]<<https://r12a.github.io/uniview/?char=1F99E>>, [Caractère Unicode non montré]<<https://r12a.github.io/uniview/?char=1F99F>>... (Si vous n'avez pas les bonnes polices, chaque lettre est un lien vers Uniview <<https://r12a.github.io/uniview/>>.)