Version 8 d'Unicode

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 19 juin 2015

https://www.bortzmeyer.org/unicode-8-0.html

Le 17 juin a vu la sortie d'une nouvelle version <http://blog.unicode.org/2015/06/announcing-unicode-s
html> du jeu de caractères Unicode, la 8.0, pile un an après la précédente <https://www.bortzmeyer.
org/unicode-7-0.html>. On peut trouver une description des principaux changements en <http:
//www.unicode.org/versions/Unicode8.0.0/> 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
--------
 120737
```

Plus de 120 000 caractères. Lesquels ont été apportés par la version 8 ?

```
ucd=> SELECT version,count(version) FROM Characters GROUP BY version ORDER BY version;
...
 7.0     |  2834
 8.0     |  7716
```

7 716 nouveaux. Lesquels ?

```
ucd=> SELECT To_U(codepoint) AS Codepoint, name FROM Characters WHERE version='8.0';
 codepoint |                                      name
-----------+-------------------------------------------------------------------------------
...
 U+8B4     | ARABIC LETTER KAF WITH DOT BELOW
...
 U+13FB    | CHEROKEE SMALL LETTER YU
...
 U+11294   | MULTANI LETTER DDHA
...
 U+12488   | CUNEIFORM SIGN DA TIMES TAK4
...
 U+1D96D   | SIGNWRITING MOVEMENT-FLOORPLANE DOUBLE ALTERNATING WRIST FLEX
...
 U+1F32F   | BURRITO
 ...
 U+1F3FF   | EMOJI MODIFIER FITZPATRICK TYPE-6
 ...
 U+1F54C   | MOSQUE
 U+1F54D   | SYNAGOGUE
...
```

Comme on le voit, c'est varié. On trouve des écritures entièrement nouvelles comme le multani ou le système Sutton pour la langue des signes, des nouvelles lettres pour des écritures existantes (le ≪ *"ARABIC LETTER KAF WITH DOT BELOW"* ≫ est ajouté à l'alphabet arabe pour écrire des langues de l'Asie du Sud-Est), des nouveaux "*emojis*" (comme le burrito)... L'alphabet cherokee a connu un grand changement, il est désormais bicaméral et les minuscules du Cherokee viennent donc s'ajouter à Unicode (comme la U+13FB ci-dessus). Par contre, comme il ne s'est écrit qu'en majuscules pendant longtemps, l'algorithme d'uniformimisation de casse d'Unicode met le Cherokee en majuscules, et pas en minuscules comme pour toutes les autres écritures.

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], (U+1F3FF a été omis, voir à la fin), [Caractère Unicode non montré], [Caractère Unicode non montré] ... (UniView `<http://r12a.github.io/uniview/>` n'a apparemment pas encore été mis à jour avec les données de la version 8).

Comme toujours depuis quelques versions d'Unicode, tout le monde s'excite sur les nouveaux "*emojis*" (si vous ne les avez pas encore sur votre ordinateur, voyez les jolis dessins sur ce site `<http://typography.guru/journal/unicode-8-emoji/>`). Plutôt que de donner des listes pittoresques, avec des glyphes rigolos, je suggère fortement à mes lecteurs de lire l'excellent "*Unicode Technical Report #51*" `<http://unicode.org/reports/tr51/>` sur la gestion des "*emojis*" dans Unicode. Vous y apprendrez le pourquoi de ces caractères, et les règles qu'ils suivent dans Unicode, et comment enregister un nouvel "*emoji*". À la lecture de ce rapport, vous saurez pourquoi le policier tue le crocodile (et pas le contraire), pourquoi un "*emoji*" dont le nom inclut "*BLACK*" n'est pas forcément noir, quelles sont toutes les formes possibles du "*shortcake*"...

Une nouveauté importante des "*emojis*" dans cette version 8 est la gestion de la couleur de peau. Traditionnellement, les personnages humains dans des "*emojis*" sont représentés avec une peau jaune vif. C'est considéré comme ≪ neutre ≫ (notez que les "*emojis*" sont nés au Japon...) Cela ne convient pas forcément à tout le monde et il y a une demande depuis longtemps pour permettre de représenter des gens à la peau rose ou noire. C'est possible avec la nouveauté des **modificateurs d'emojis**. Ce sont

1. Car trop difficile à faire afficher par LaTeX

des caractères Unicode qui se placent après l'"*emoji*" et modifient la couleur de la peau du personnage. Ils sont cinq, correspondant aux cinq degrés de l'échelle de Fitzpatrick <http://www.arpansa.gov.au/pubs/RadiationProtection/FitzpatrickSkinType.pdf>. Par exemple, si votre navigateur gère ces récents modificateurs, en mettant les deux caractères U+1F477 ("*CONSTRUCTION WORKER*") et U+1F3FE ("*EMOJI MODIFIER FITZPATRICK TYPE-5*"), vous devriez voir un ouvrier à la peau sombre (mais pas complètement noire). Essayons d'abord sans le modificateur : [Caractère Unicode non montré] et avec : [Caractère Unicode non montré] [Caractère Unicode non montré]. (Au passage, ce caractère U+1F477 est cité dans le rapport #51 pour une autre raison politique : il est neutre du point de vue du genre et peut donc être représenté par un homme ou une femme.)

À noter deux limites des modificateurs : ils ne peuvent pas s'appliquer à chaque membre d'un groupe. Donc, dans U+1F46C "*TWO MEN HOLDING HANDS*", les deux hommes ont forcément la même couleur de peau. Et, autre limite, les modificateurs ne peuvent changer que la couleur de la peau, pas d'autres caractéristiques comme la minceur.