Version 14 d'Unicode

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 16 septembre 2021

https://www.bortzmeyer.org/unicode-14-0.html

Le mardi 14 septembre est sortie la version 14 d'Unicode <http://blog.unicode.org/2021/09/announcing-unicode-standard-version-140.html> (retardée pour cause de Covid-19 <http://blog.unicode.org/2020/04/unicode-140-delayed-for-6-months.html>). Une description officielle des principaux changements est disponible <https://www.unicode.org/versions/Unicode14.0.0/> 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
--------
 144762
```

Combien de caractères sont arrivés avec la version 14 ?

```
ucd=> SELECT version,count(version) FROM Characters GROUP BY version ORDER BY version::float;
...
 11.0    |   684
 12.0    |   554
 12.1    |     1
 13.0    |  5930
 14.0    |   838
```

838 nouveaux caractères, c'est une version calme (la version 13 <https://www.bortzmeyer.org/unicode-13-0.html> apportait bien davantage de caractères). Quels sont ces nouveaux caractères ?

```
ucd=> SELECT To_U(codepoint) AS Code_point, name FROM Characters WHERE version='14.0' ORDER BY Codepoint;
 code_point |                                    name
------------+-------------------------------------------------------------------------
...
 U+870      | ARABIC LETTER ALEF WITH ATTACHED FATHA
 U+871      | ARABIC LETTER ALEF WITH ATTACHED TOP RIGHT FATHA
 U+872      | ARABIC LETTER ALEF WITH RIGHT MIDDLE STROKE
 U+873      | ARABIC LETTER ALEF WITH LEFT MIDDLE STROKE
 U+874      | ARABIC LETTER ALEF WITH ATTACHED KASRA
...
 U+12F90    | CYPRO-MINOAN SIGN CM001
 U+12F91    | CYPRO-MINOAN SIGN CM002
 U+12F92    | CYPRO-MINOAN SIGN CM004
 U+12F93    | CYPRO-MINOAN SIGN CM005
 ...
 U+1CF00    | ZNAMENNY COMBINING MARK GORAZDO NIZKO S KRYZHEM ON LEFT
 U+1CF01    | ZNAMENNY COMBINING MARK NIZKO S KRYZHEM ON LEFT
 U+1CF02    | ZNAMENNY COMBINING MARK TSATA ON LEFT
...
 U+1E290    | TOTO LETTER PA
 U+1E291    | TOTO LETTER BA
 U+1E292    | TOTO LETTER TA
 U+1E293    | TOTO LETTER DA
...
 U+1FAAC    | HAMSA
 U+1FAB7    | LOTUS
 U+1FAB8    | CORAL
 U+1FAB9    | EMPTY NEST
 U+1FABA    | NEST WITH EGGS
 ...
```

Cette version amène en effet plusieurs écritures nouvelles comme le toto, le chypro-minoen (toujours pas déchiffré, ce qui explique que ses caractères aient un numéro et pas un nom), le vieil ouïghour (écriture pour laquelle ont été fabriquées les plus anciens caractères mobiles <https://fr.wikipedia.org/wiki/Alphabet_ou%C3%AFghour#/media/Fichier:Beijing_printing_museum.Caract%C3%A8res_mobiles_en_ancien_Ouighour.jpg> connus), le vithkuqi, ou le tangsa. On trouve également beaucoup de nouveaux caractères de l'alphabet arabe pour écrire des langues non-arabes.

Plus anecdotique, on trouve également la notation musicale Znamenny.

Si vous avez les bonnes polices de caractères, vous allez pouvoir voir quelques exemples (sinon, le lien mène vers Uniview <https://r12a.github.io/uniview/>). Un nouveau symbole apparait, le Som ([Caractère Unicode non montré [1]] <https://r12a.github.io/uniview/?char=20C0>), pour la monnaie officielle du Kirghizistan. Et on a l'habituelle succession <https://www.unicode.org/emoji/charts-14.0/emoji-released.html> de nouveaux émojis, plus ou moins utiles. Les recruteurs aimeront l'index pointé [Caractère Unicode non montré] <https://r12a.github.io/uniview/?char=1FAF5>, les royalistes seront contents de la tête couronnée [Caractère Unicode non montré] <https://r12a.github.io/uniview/?char=1FAC5>, les nostalgiques d'Usenet auront le troll [Caractère Unicode non montré] <https://r12a.github.io/uniview/?char=1F9CC>, les fanas de Dune frémiront en voyant l'eau qu'on verse [Caractère Unicode non montré] <https://r12a.github.io/uniview/?char=1FAD7>, les "*geeks*" s'angoisseront de la batterie vide [Caractère Unicode non montré] <https://r12a.github.io/uniview/?char=1FAAB> et les partisans de la surveillance de la population noteront qu'on a un émoji « carte d'identité » [Caractère Unicode non montré] <https://r12a.github.io/uniview/?char=1FAAA>. Les personnes transgenres remarqueront l'homme enceint ([Caractère Unicode non montré] <https://r12a.github.

1. Car trop difficile à faire afficher par LaTeX

io/uniview/?char=1FAC3>, ainsi que la personne enceinte [Caractère Unicode non montré] <https://r12a.github.io/uniview/?char=1FAC4>, la femme enceinte [Caractère Unicode non montré] <https://r12a.github.io/uniview/?char=1F930> était arrivée en Unicode version 9).

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
-------
  7741
```

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
------------------+-------
 Modifier_Symbol |   125
 Other_Symbol    |  6605
 Math_Symbol     |   948
 Currency_Symbol |    63
```