

# Naviguer dans les caractères Unicode

Stéphane Bortzmeyer

<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 27 Février 2008

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

---

Le jeu de caractères Unicode est vaste (99 000 caractères la dernière fois que j'ai compté et cela augmente régulièrement). Il est donc souvent difficile de trouver un caractère donné. Quels sont les outils pour cela ?

Prenons l'exemple des caractères permettant une citation comme les guillemets en français. Cela se dit "*quotation*" dans la langue de Donald Westlake. Peut-on trouver tous les caractères de citation facilement ?

Mon outil en ligne préféré, utilisant toutes les techniques modernes du Web est Uniview <<http://people.w3.org/rishida/scripts/uniview>>. Taper "*quotation*" dans le moteur de recherche permet d'avoir la liste et leur apparence (si la bonne police est bien connue du navigateur).

Autre outil en ligne, moins convivial mais plus adapté pour des extractions de masse du genre « Tous les caractères de ponctuation qui sont identiques à leur forme canonique » : Unicode Utilities <<http://unicode.org/cldr/utility/>>. Ainsi, on peut extraire facilement, par exemple, tous les caractères introduits entre les version 4.1 et 6.0 incluses <[http://unicode.org/cldr/utility/list-unicodeset.jsp?a=\[\p{age%3D6.0}-\p{age%3D4.1}\]\&g=age](http://unicode.org/cldr/utility/list-unicodeset.jsp?a=[\p{age%3D6.0}-\p{age%3D4.1}]\&g=age)>.

Si on veut simplement afficher tous les caractères Unicode (dans les limites de son navigateur Web), je recommande Unicodinator <<http://unicodinator.com/>>. Si on connaît le point de code, <<http://www.fileformat.info/info/unicode/index.htm>> permet, entre autre chose, d'y accéder directement via un URL qui contient ce point de code. Regardez <<http://www.fileformat.info/info/unicode/char/1f304/index.htm>>, <<http://www.fileformat.info/info/unicode/char/e9/index.htm>> ou <<http://www.fileformat.info/info/unicode/char/fe94/index.htm>>.

On peut aussi vouloir chercher dans les documents officiels et le consortium Unicode a une excellente page pour cela <<http://www.unicode.org/standard/where/>>.

Pour le cas où on connaît la forme du caractère et où on cherche son nom et son point de code, l'excellent reconnaisseur de formes `<http://www.shapecatcher.com/>` peut aider.

Et si on préfère travailler en local, sans nécessiter un accès à Internet, et avec ses propres outils ?

Pour ceux qui aiment Emacs, le plus simple est certainement de charger le fichier de la norme (sur une Debian, il est dans `/usr/share/unicode/UnicodeData.txt` si on a installé le paquetage `unicode-data` - voici l'intérêt des normes gratuites `<http://www.bortzmeyer.org/itu-normes-gratuites.html>`, leurs fichiers peuvent être inclus dans un système de paquetages). On peut alors utiliser la fonction de recherche d'Emacs, `C-s` et c'est parti :

```
00AB;LEFT-POINTING DOUBLE ANGLE QUOTATION MARK;Pi;0;ON;;;;;Y;LEFT POINTING GUILLEMET;*;;;
...
00BB;RIGHT-POINTING DOUBLE ANGLE QUOTATION MARK;Pf;0;ON;;;;;Y;RIGHT POINTING GUILLEMET;*;;;
```

Pour ceux qui préfèrent SQL, une approche possible est décrite dans mon article `La base de données Unicode en SQL` `<http://www.bortzmeyer.org/unicode-to-sql.html>`.

```
ucd=> SELECT To_U(codepoint) AS U_Codepoint, name FROM Characters WHERE name LIKE '%QUOTATION%' ORDER BY codepoint
u_codepoint | name
-----+-----
U+22        | QUOTATION MARK
U+AB        | LEFT-POINTING DOUBLE ANGLE QUOTATION MARK
U+BB        | RIGHT-POINTING DOUBLE ANGLE QUOTATION MARK
U+2018      | LEFT SINGLE QUOTATION MARK
```

Enfin, pour les Vrais Hommes `<http://www.bortzmeyer.org/command-line.html>` qui préfèrent la Ligne de Commande, l'excellent programme `unicode`, écrit pour Debian (mais qui tourne sans doute sur d'autres Unix) :

```
# Available at http://packages.debian.org/etch/unicode
% unicode quotation
U+0022 QUOTATION MARK
UTF-8: 22 UTF-16BE: 0022 Decimal: &#34;
"
Category: Po (Punctuation, Other)
Bidi: ON (Other Neutrals)

U+00AB LEFT-POINTING DOUBLE ANGLE QUOTATION MARK
UTF-8: c2 ab UTF-16BE: 00ab Decimal: &#171;
...
```