# RFC 7997 : The Use of Non-ASCII Characters in RFCs

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 16 décembre 2016

Date de publication du RFC : Décembre 2016

http://www.bortzmeyer.org/7997.html

---

Les RFC sont forcément écrits en anglais, qui restera la langue officielle (cf. RFC 7322[1]). L'anglais peut s'écrire avec uniquement les caractères ASCII (avec quelques exceptions : "*resume*" <https://en.wiktionary.org/wiki/resume> et "*résumé*" <https://en.wiktionary.org/wiki/r%C3%A9sum%C3%A9> ne sont pas le même mot). Mais on pourra désormais inclure des caractères non-ASCII, par exemple pour le nom des auteurs (chic, je pourrais écrire correctement mon prénom dans les RFC). Cette possibilité permettra aussi aux exemples de protocoles Internet utilisant Unicode (la grande majorité) d'être plus lisibles.

Cette nouvelle possibilité fait partie de celles qu'offre le nouveau format des RFC, décrit dans le RFC 7990. Il n'y a quand même pas d'autorisation générale d'inclure n'importe quel caractère Unicode dans les RFC, à n'importe quel endroit. Le "*RFC Editor*" pourra toujours refuser tel ou tel caractère, par exemple parce qu'il n'existe pas de police permettant de l'afficher. Et le « non-ASCII » n'est autorisé que dans certains cas, décrits plus loin. La grande majorité du texte sera donc du pur ASCII (RFC 20).

L'encodage de ces caractères sera bien sûr UTF-8.

Il ne suffit pas de proclamer « on a droit à Unicode ». Il faut aussi adapter les outils. Par exemple, notre RFC impose (section 2) que les outils de recherche dans les RFC gèrent correctement la recherche Unicode. (C'est pour traiter le cas des outils imparfaits que le RFC demande aussi que les noms d'auteurs en Unicode soient accompagnés d'une version en ASCII.) Et que le RFC soit affichable correctement sur un bon nombre de plate-formes (d'où la possibilité de rejeter les caractères les plus rares).

Ce problème du repli (vers une version en ACSII pur) est souvent cité dans le RFC. Ainsi, lorsqu'on veut mentionner un caractère Unicode (mettons le thorn islandais), le RFC permet désormais

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc7322.txt

de l'afficher proprement, mais il demande qu'on l'accompagne du numéro du point de code, et, si possible, de son nom Unicode. Cela donnerait, par exemple « *"For instance, U+00FE, "LATIN SMALL LETTER THORN", [Caractère Unicode non montré* [2] *], is interesting because...* » ". Notez que cette façon de désigner des caractères Unicode que tout le monde n'arrivera pas forcément à afficher n'est pas vraiment standardisée. Dans les RFC actuels, on trouve des variantes (voir cette discussion <https://www.rfc-editor.org/pipermail/rfc-interest/2016-August/009716.html>). Le RFC contient plusieurs exemples sur la façon d'écrire la phrase « *"Temperature changes in the Temperature Control Protocol are indicated by the U+2206 character ([Caractère Unicode non montré ], "INCREMENT")"* », tous acceptés (le nom Unicode n'est pas obligatoire, il peut être placé avant ou après le caractère lui-même, etc.) Autre cas, ce texte du RFC 8264, « *"For example, the characters U+13DA U+13A2 U+13B5 U+13AC U+13A2 U+13AC U+13D2 from the Cherokee block look similar to the ASCII characters "STPETER""* » deviendrait « *"For example, the characters U+13DA U+13A2 U+13B5 U+13AC U+13A2 U+13AC U+13D2 ([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 ]) from the Cherokee block look similar to the ASCII characters "STPETER""* ». Des tables comme celles des identificateurs et mots de passe Unicode légaux (RFC 8265) seraient ainsi bien plus lisibles.

Pour les noms, par exemple ceux des auteurs. On aurait du « non-ASCII » et un texte de repli, comme (en utilisant le vocabulaire XML du RFC 7991) :

```
<author fullname="&#1512;&#1493;&#1504;&#1497; &#1488;&#1489;&#1503;" asciiFullname="R. Even"/>
<author fullname="&#21556;&#38054;" asciiFullname="Q. Wu"/>
<author fullname="J. Smith" asciiFullname="J. Smith"/> <!-- Oui, dans
ce cas, il faut le dire deux fois -->
```

Cela permettra enfin d'écrire correctement les noms des auteurs de RFC.

La bibliographie d'un RFC est également un bon endroit où mettre des caractères Unicode, par exemple lorsqu'on cite des textes non-anglo-saxons. Ainsi, la bibliographie du RFC 5933 pourrait inclure :

```
[GOST3410] "Information technology. Cryptographic data security.
           Signature and verification processes of [electronic]
           digital signature.", GOST R 34.10-2001, Gosudarstvennyi
           Standard of Russian Federation, Government Committee of
           Russia for Standards, 2001. (In Russian)

           "&#1048;&#1085;&#1092;&#1086;&#1088;&#1084;&#1072;&#1094;&#1080;&#1086;&#1085;&#1085;&#1072;&#110
           &#1080;&#1085;&#1092;&#1086;&#1088;&#1084;&#1072;&#1094;&#1080;&#1080;. &#1055;&#1088;&#1086;&#10
           &#1101;&#1083;&#1077;&#1082;&#1090;&#1088;&#1086;&#1085;&#1085;&#1086;&#1081; &#1094;&#1080;&#109
           &#1043;&#1086;&#1089;&#1091;&#1076;&#1072;&#1088;&#1089;&#1090;&#1074;&#1077;&#1085;&#1085;&#1099
```

Le second texte étant l'original russe.

Les règles exactes figurent dans la section 3. D'abord, on peut mettre du « non-ASCII » comme on veut quand il fait partie d'un exemple. Ainsi, la communication XMPP pourrait être décrite de manière plus naturelle. Au lieu de cet exemple de communication en tchèque (RFC 6121) :

2. Car trop difficile à faire afficher par LaTeX

```
<message
    from='juliet@example.com/balcony'
    id='z94nb37h'
    to='romeo@example.net'
    type='chat'
    xml:lang='en'>
  <body>Wherefore art thou, Romeo?</body>
  <body xml:lang='cs'>
     Pro&#x010D;e&#x017D; jsi ty, Romeo?
   </body>
</message>
```

On pourra écrire la forme lisible :

```
<message
    from='juliet@example.com/balcony'
    id='z94nb37h'
    to='romeo@example.net'
    type='chat'
    xml:lang='en'>
  <body>Wherefore art thou, Romeo?</body>
  <body xml:lang='cs'>
     Pro&#269;e&#381; jsi ty, Romeo?
   </body>
</message>
```

Ensuite, on peut utiliser le « non-ASCII » pour les cas cités plus haut (noms d'auteurs, textes non-anglophones dans la bibliographie, etc). Pour les exemples utilisant un langage de programmation, notre RFC spécifie qu'il faut suivre les règles du langage en question. Ainsi, Python 3 autorisant l'Unicode même dans les noms de variables, on peut écrire :

```
a = "chocolat"
b = "café"       # Accentué
ç = "lait"
print(a+b+ç)
```

Enfin, un petit mot sur la normalisation Unicode, pour rappeler que le format des RFC ne garantit rien à ce sujet (on aurait pu décider que NFC serait systématiquement utilisée...) et que les auteurs de RFC ne doivent donc pas compter dessus.

Le premier RFC publié avec des caractères Unicode a été le RFC 8187, en septembre 2017.