

RFC 5198 : Unicode Format for Network Interchange

Stéphane Bortzmeyer

<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 30 Mars 2008

Date de publication du RFC : Mars 2008

<http://www.bortzmeyer.org/5198.html>

Il ne reste plus beaucoup de normes Internet qui soient limitées à l'ASCII et elles tombent une à une. Un des problèmes rencontrés pour éliminer les dernières est que certains protocoles transmettent du **texte** et que « texte », à l'IETF, voulait toujours dire seulement ASCII. Voici que ce RFC introduit un nouveau type de texte, utilisant Unicode.

Pendant longtemps, plusieurs normes comme le RFC 959¹ sur FTP faisaient référence à des données de type **texte**, c'est-à-dire sans formatage, avec des lignes terminées par la séquence CRLF ("*Carriage Return*" puis "*Line Feed*"). Cette définition du texte, souvent nommé « NVT », datant du RFC 854, limitait le texte au jeu de caractères ASCII, jeu de caractères qui ne permet pas de représenter toutes les écritures du monde. Où est le problème ? peut-on se demander. Pourquoi ne pas simplement remplacer ASCII par Unicode partout dans le texte ? C'est en effet la bonne solution mais le diable est dans les détails et, Unicode étant beaucoup plus riche qu'ASCII et n'étant pas figé, traiter ces détails a pris du temps. (Et ce RFC a fait l'objet de nombreuses controverses.) Unicode a plusieurs encodages, plusieurs façons de représenter une fin de ligne et même plusieurs façons de représenter le même caractère.

Le RFC pose donc comme règles (section 2) :

- Jeu de caractères Unicode.
- Utilisation de l'encodage UTF-8, normalisé dans le RFC 3629.
- Normalisation obligatoire avec le mécanisme NFC, qui combine le plus possible les caractères.
- Utilisation de la séquence CRLF pour les sauts de ligne.

¹Pour voir le RFC de numéro NNN, <http://www.ietf.org/rfc/rfcNNN.txt>, par exemple <http://www.ietf.org/rfc/rfc959.txt>

Le choix d'UTF-8 provient de sa large utilisation dans beaucoup de protocoles Internet, due en partie au fait qu'il est un sur-ensemble de l'encodage ASCII.

Le choix du CRLF comme indicateur d'une fin de ligne est également dû à sa large utilisation aujourd'hui. Bien qu'Unicode aie des caractères plus adaptés ("*Line separator*", U+0028 et "*Paragraph separator*", U+0029), ils sont très peu utilisés. La passionnante annexe C discute en détail ce choix.

La normalisation fait l'objet d'une section 3 détaillée. Elle est nécessaire car Unicode permet de représenter un caractère de plusieurs façons. Ainsi, U+2126, "*Ohm sign*", [Caractère Unicode non montré²], a une forme canonique qui est U+03A9, "*Greek capital letter omega*", [Caractère Unicode non montré], (les symboles scientifiques ne sont normalement pas codés séparément dans Unicode et ceux qui le sont sont normalisés dans une lettre ordinaire). Cas plus connu, deux caractères peuvent se **combiner** par exemple U+0061 U+0300 se combinent, par le mécanisme NFC, en U+00E0 (la première séquence est faite d'un petit a et de l'accent grave combinant, la seconde est donc juste un à). L'obligation de normalisation rendra donc plus facile la comparaison des textes (la section 6 sur la sécurité note toutefois que le pare-feu prudent ne doit pas supposer que l'expéditeur respectera la norme).

La section 4 couvre un autre problème délicat, celui des versions d'Unicode. Contrairement à ASCII, qui était figé dès le début, des nouveaux caractères sont ajoutés dans Unicode régulièrement, ainsi que de nouvelles règles de normalisation. Si on ne fait pas attention, une nouvelle version d'Unicode pourrait rendre invalides (car non normalisés) des textes auparavant valables. Pour éviter cela, notre RFC interdit l'utilisation des points de code non affectés. Les politiques de stabilité du consortium Unicode garantissent que, dans ce cas, la normalisation NFC est stable (la section 5.2 discute également ce choix).

La section 5 explique quelles normes pourraient tirer profit de ce nouveau RFC. Celles qui utilisent déjà UTF-8, comme LDAP, ne sont pas concernées. Mais, outre FTP, déjà cité, cette nouvelle référence pourrait intéresser des protocoles comme whois (dont le RFC 3912 note que, en théorie, il est limité à ASCII).

Cas rare chez les RFC, il se termine par des annexes traitant d'histoire. L'excellente annexe A revient sur l'histoire tourmentée des jeux de caractères dans Internet, du RFC 20, qui imposera ASCII, au RFC 318 qui sera le premier à mentionner le « NVT » ("*Network Virtual Terminal*"). L'annexe B, elle, extrait de différents RFC, ce qui est à ce jour la définition la plus complète et la plus correcte de la norme NVT... que notre RFC remplace. Bel hommage du présent au passé.

²Car trop difficile à faire afficher par L^AT_EX