

RFC 7994 : Requirements for Plain-Text RFCs

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 16 décembre 2016

Date de publication du RFC : Décembre 2016

<https://www.bortzmeyer.org/7994.html>

Dans la grande série des nombreux RFC spécifiant le nouveau format de ces documents (avec XML étant désormais la référence), ce court RFC décrit le format de publication « texte brut » des RFC.

En effet, si le format « texte brut » n'est plus la référence <<http://www.rfc-editor.org/pipermail/rfc-interest/2013-May/005584.html>> des RFC, ce format reste toujours utile. Dans le nouveau mécanisme (RFC 7990¹), il y a désormais plusieurs formats de **publication**, obtenus à partir du format canonique (celui qui est en XML). Le texte brut est l'un d'eux. Ce format, historiquement le seul utilisé par les RFC, ne disparaît pas, mais perd de son importance. Il n'est plus qu'un des formats de publication parmi d'autres (comme HTML, voir le RFC 7992). Le RFC 6949 expliquait ce choix.

À noter que les RFC produits avant ce changement ne sont pas affectés : leur format de référence reste le texte brut (et en ASCII seul).

Bon, désormais, on aura un format de sortie (parmi d'autres) en « texte seul » ou « texte brut ». Mais c'est quoi, le texte brut? Notre RFC reprend la définition du consortium Unicode : « du texte encodé pour les ordinateurs composé uniquement de points de code d'une norme donnée, sans instructions de format ou de structure ». Bref, les caractères indiqués ne valent que pour eux-mêmes, ils n'indiquent jamais de formatage ou de style. Selon cette définition, HTML, LaTeX et Markdown (RFC 7763) ne sont donc pas du texte brut. (La définition n'est pas 100 % parfaite. La norme Unicode, par exemple, inclut des caractères qui influencent le format.) Le texte brut est donc ce qui est le plus portable : tous les acteurs qui connaissent la norme de jeu de caractères sous-jacente (aujourd'hui, quasiment toujours Unicode) peuvent lire et écrire du texte brut. C'est d'ailleurs une des raisons pour lesquelles les RFC ont si longtemps gardé ce format <<https://www.bortzmeyer.org/rfc-en-texte-brut.html>> comme format canonique.

1. Pour voir le RFC de numéro NNN, <https://www.ietf.org/rfc/rfcNNN.txt>, par exemple <https://www.ietf.org/rfc/rfc7990.txt>

Mais si le texte brut n'est pas idéal comme format de référence, il reste un format de sortie très utile, notamment pour son interopérabilité, ou en raison de l'existence de nombreux outils qui peuvent le traiter (à commencer par `grep`...) Désormais, donc, le format canonique est le XML décrit dans le RFC 7991 et le texte brut sera produit automatiquement par les nouveaux outils. Mais ce texte brut a des règles qui sont légèrement différentes du texte brut original (« RFC canal historique ») et notre RFC 7994 les décrit. Il est très court, car le format « texte brut » est un format simple.

D'abord, le jeu de caractères (section 2). Ce sera bien sûr Unicode, mais avec les restrictions indiquées dans le RFC 7997. En pratique, là où les caractères non-ASCII ne sont pas autorisés, il faudra utiliser l'ASCII équivalent, donné dans les attributs XML prévus à cet effet (`ascii`, RFC 7991 en section 2.23.1, `asciiFullname` en 2.7.1, etc). L'encodage sera obligatoirement UTF-8 (RFC 3629). Curieusement, il est prévu de mettre une BOM au début du document.

Que faire avec les graphiques, désormais autorisés par le RFC 7990, et écrits en SVG (RFC 7996)? Ces graphiques sont, dans la source XML, à l'intérieur d'un élément `<artwork>`. Comment les rendre dans du texte brut (section 3 de notre RFC)? D'abord, si le graphique n'est pas en SVG mais dans le traditionnel art ASCII (indiqué par `type=ascii-art`), on utilise cet art ASCII. Autrement, notre RFC ne propose pas de solution générale. Il est recommandé aux auteurs de diagrammes et schémas de prévoir une alternative en art ASCII, même quand ils font du SVG.

Enfin, la section 4 du RFC couvre le problème de la « mise en page ». Un caractère de fin de page (U+000C) sera inclus automatiquement toutes les 58 lignes (les outils auront probablement une option pour ne pas inclure de telles marques). L'outil devra gérer le délicat problème des veuves et des orphelins. Les lignes feront 72 caractères, suivies de deux caractères marquant la fin (U+000D U+000A).

Les textes de début du RFC (RFC 5741) seront automatiquement mis comme avant, mais les en-têtes et pieds de page disparaissent. Autre disparition, il n'y aura plus, dans le format de sortie en texte brut, de numéros de pages dans la table des matières (dommage, je trouve, mais c'est au nom de la cohérence avec les autres formats de sortie).