RFC 3454 : Preparation of Internationalized Strings ("stringprep")

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 13 novembre 2006

Date de publication du RFC : Décembre 2002

https://www.bortzmeyer.org/3454.html

Beaucoup de protocoles ont besoin de manipuler des chaînes de caractères Unicode. La grande taille du répertoire Unicode fait qu'il est relativement fréquent de rencontrer deux chaînes différentes selon Unicode, mais identiques selon les utilisateurs. Il est donc nécessaire de définir des fonctions de normalisation. Ce que fait notre RFC. (Il a ensuite été remplacé par une méthode assez différente, décrite dans le RFC 7564 [1].)

Avec les petits jeux de caractères comme US-ASCII, tout est simple. Deux chaînes différentes selon ASCII le sont également selon les utilisateurs humains. Le seul cas où une certaine normalisation est nécessaire est celui de l'insensibilité à la casse, lorsqu'on décide que (en prenant un exemple DNS) `wikipedia.fr` et `WIKIpediA.Fr` sont équivalents (ont la même forme canonique).

Avec Unicode, ce n'est plus le cas : par exemple, la chaîne composée de l'unique caractère U+00E8 ("*LATIN SMALL LETTER E WITH GRAVE*") est différent de la chaîne formée des caractères U+0065 ("*LATIN SMALL LETTER E*") et U+0300 ("*COMBINING GRAVE ACCENT*"), alors que, pour la grande majorité des applications, la première (dite « précomposée ») est certainement « identique » à la seconde.

De même, en allemand, le caractère U+00DF ("*LATIN SMALL LETTER SHARP S*" ou [Caractère Unicode non montré [2]]) est souvent considéré comme « identique » à la chaîne "ss".

1. Pour voir le RFC de numéro NNN, `https://www.ietf.org/rfc/rfcNNN.txt`, par exemple `https://www.ietf.org/rfc/rfc7564.txt`
2. Car trop difficile à faire afficher par LaTeX

Les protocoles qui comparent ou classent des chaînes de caractères Unicode doivent donc **normaliser** ces chaînes avant toute comparaison. Aucune normalisation ne convient à tous les cas et notre RFC ne spécifie donc qu'un cadre général, qui doit être décliné en **profils** selon les besoins de l'application.

Notre RFC normalise donc ce cadre, les tables à utiliser et les points à spécifier dans chaque profil. Ensuite, d'autres RFC font définir les profils. Par exemple, le RFC 3491 normalise le profil **Nameprep**, utilisé autrefois par les noms de domaines en Unicode (IDN). Selon Nameprep, `Strasse.de` est ainsi identique à `stra[Caractère Unicode non montré ]e.de`. Plusieurs RFC spécifiaient un tel profil, comme le RFC 3491 déjà cité ou bien comme le RFC 4013, qui normalisait **SASLprep**, utilisé pour normaliser les noms lors d'opérations d'authentification (depuis remplacé par le RFC 7613) ou comme le RFC 6122 qui normalisait les adresses XMPP (depuis remplacé par une autre norme sans Stringprep, le RFC 7622) et les profils **resourceprep** ou **nodeprep**.

Stringprep n'est plus utilisé aujourd'hui, et la méthode pour gérer les chaînes de caractères en Unicode est désormais celle du RFC 7564.