RFC 7790 : Mapping characters for PRECIS classes

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 1 mars 2016

Date de publication du RFC : Février 2016

https://www.bortzmeyer.org/7790.html

PRECIS ("*preparation, enforcement, and comparison of internationalized strings*"), décrit désormais dans le RFC 8264 [1], est un cadre général pour l'internationalisation d'identificateurs sur l'Internet. Il expose un certain nombre de règles, parmi lesquelles les développeurs de protocoles Internet et d'applications vont pouvoir choisir, au lieu de partir de zéro à chaque fois. Ce nouveau et court RFC donne des indications aux concepteurs de nouveaux profils PRECIS sur les correspondances ("*mappings*") à effectuer (par exemple, mais pas uniquement, entre des caractères majuscules et minuscules). Il n'est pas normatif : l'internationalisation est quelque chose de compliqué et il n'est pas facile d'obtenir un consensus. L'utilisateur indique un identificateur, en quoi peut-on le transformer ?

Au départ de cet identificateur, on trouve une saisie par l'utilisateur (clavier physique ou clavier virtuel), ou bien une sélection par ce même utilisateur (copier/coller depuis du texte, choix d'un signet, choix parmi les résultats d'un moteur de recherche...) À ce stade, l'utilisateur ne tient pas compte de certaines caractéristiques du texte, qu'il considère non pertinentes. C'est ainsi qu'il va ignorer la casse, ou bien la largeur du caractère (pour les caractères qui ont une version étroite et une version large comme le point d'exclamation, qui existe en U+0021, « ! » et U+FF01, « [Caractère Unicode non montré [2]] ») Il est donc souhaitable d'avoir une phase de **correspondance** ("*mapping*") entre cette saisie ou cette sélection et le moment où l'identificateur est passé au programme. C'est lors que cette phase qu'on va, par exemple, tout mettre en minuscules, et remplacer les caractères larges par leur version étroite.

Pour le cas des noms de domaine (IDN), cette correspondance est déjà traitée dans le RFC 5895. Pour les autres identificateurs, le cadre PRECIS ("*PReparation, Enforcement, and Comparison of Internationalized Strings*"), dans le RFC 8264, s'en charge. Mais il ne traite qu'une partie des correspondances souhaitables (comme la casse) et il était donc souhaitable de l'étendre. C'est ce que fait ce nouveau RFC.

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc8264.txt
2. Car trop difficile à faire afficher par LaTeX

La section 2 de notre court RFC spécifie ces correspondances supplémentaires. Par exemple, les délimiteurs (comme le point). Chaque protocole et format a ses propres délimiteurs (@ pour le courrier électronique, / dans les URL, etc). L'idée est de faire correspondre les caractères qui **ressemblent** au(x) délimiteur(s) du protocole vers le caractère canonique. Ainsi, si un protocole utilise le point (U+002E) come délimiteur, et que le texte saisi par l'utilisateur contient un point idéographique (U+3002, alias « [Caractère Unicode non montré] »), il est logique de remplacer ces points par le point « officiel » du protocole.

Après la correspondance des délimiteurs, les « correspondances spéciales ». Le terme désigne des règles qui sont différentes des règles par défaut de PRECIS. Par exemple, supprimer les caractères de contrôle (pas les transformer en un autre caractère, non, les supprimer franchement) ou bien transformer des espaces différents du caractère espace d'ASCII en espaces ASCII (U+0020). C'est ce que font des protocoles comme EAP (RFC 3748) ou les ACL d'IMAP (RFC 4314).

Autre cas rigolo, celui où le changement de casse dépend de la "*locale*", c'est-à-dire en général de la langue utilisée. L'exemple le plus fameux, toujours mentionné dans les discussions Unicode, vient du turc où la minuscule du I (U+0049) n'est pas i (U+0069) mais [Caractère Unicode non montré] (U+0131, et encore, la règle exacte est un peu plus compliquée). Comme PRECIS est conçu pour des protocoles Internet, qui ne connaissent pas la langue, les applications (qui, elles, connaissent souvent la langue de leur utilisateur) doivent traiter ce cas, avant de passer l'information sur l'Internet (voir aussi l'annexe B et C, pour des cas spéciaux dans le cas spécial).

L'annexe A de notre RFC résume dans un tableau ces correspondances selon divers protocoles (IDN, iSCSI, EAP, etc).