RFC 6452 : The Unicode code points and IDNA - Unicode 6.0

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 13 novembre 2011

Date de publication du RFC : Novembre 2011

https://www.bortzmeyer.org/6452.html

Une des exigences de la norme sur les IDN (RFC 5890 [1]) est la **stabilité** : si un caractère est autorisé dans les noms de domaine à un moment donné, il devrait l'être pour toujours, sous peine qu'une épée de Damoclès pèse sur les titulaires de noms, la crainte qu'un nom qu'ils utilisent devienne subitement invalide. Mais le mécanisme par lequel sont déterminés les caractères autorisés ou interdits rend inévitable de tels problèmes, heureusement dans des cas rares et marginaux. C'est ce qui vient de se produire avec la sortie de la norme Unicode version 6 <https://www.bortzmeyer.org/unicode-6-0.html> : le caractère [Caractère Unicode non montré [2]] (U+19DA), autrefois autorisé, est désormais interdit. S'il y avait eu des noms de domaine l'utilisant (ce n'était apparemment pas le cas), ils deviendraient subitement illégaux.

Vu que ce caractère est très rare, toute cette discussion n'est-elle pas une tempête dans un verre d'eau ? Pas complètement, parce que c'est la première fois que le problème se pose depuis la sortie de la version 2 de la norme IDN <https://www.bortzmeyer.org/idnabis.html>, et que la façon dont ce cas a été traité va servir de modèle pour des futurs problèmes qui pourraient être plus délicats.

Reprenons dès le début : dans la norme actuelle sur les IDN, dite « IDNAbis » (RFC 5890), les caractères Unicode peuvent être valides ou invalides dans un nom de domaine. Cette validité (qui était déterminée manuellement dans la première version d'IDN) découle d'un algorithme, normalisé dans le RFC 5892, qui prend en compte les propriétés que la norme Unicode attache à un caractère. Par exemple, si le caractère a la propriété « Est une lettre », il est valide. Ce système a l'avantage de remédier au principal inconvénient d'IDN version 1, le fait que le premier IDN était lié à une version particulière d'Unicode. Avec IDNAbis, au contraire, lorsque le consortium Unicode sort une nouvelle version de sa

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc5890.txt
2. Car trop difficile à faire afficher par LaTeX

norme, il suffit de faire tourner à nouveau l'algorithme, et on a la liste des caractères valides dans cette nouvelle version.

Le problème que cela pose est que les propriétés d'un caractère Unicode peuvent changer d'une version à l'autre. Elles ne sont pas forcément couvertes par les engagements de stabilité du consortium Unicode. C'est ce qui s'est produit en octobre 2010, lors de la sortie de la version 6 <https://www.bortzmeyer.org/unicode-6-0.html> d'Unicode. Trois caractères ont vu leur validité changer car leur "*GeneralCategory*" Unicode a changé. Pour les deux premiers, rien de très grave. [Caractère Unicode non montré] (U+0CF1) et [Caractère Unicode non montré] (U+0CF2), classés à tort comme des symboles, ont été reclassés comme lettres et sont donc devenus valides, alors qu'ils étaient invalides. Personne n'avait donc pu les utiliser et le changement ne va donc pas affecter les titulaires de noms de domaine.

Mais le troisième pose un problème épineux. Pas quantitativement. Ce caractère [Caractère Unicode non montré] (U+19DA), de l'écriture Tai Lue n'a apparemment jamais été utilisé dans un nom de domaine, en partie par ce que l'écriture dont il fait partie est peu répandue (et ses utilisateurs sont peu connectés à l'Internet). Mais, qualitativement, c'est ennuyeux car U+19DA (un chiffre) fait le chemin inverse. D'autorisé, il devient interdit. Potentiellement, il peut donc remettre en cause la stabilité des noms de domaine. Si on applique aveuglément l'algorithme du RFC 5892, U+19DA va devenir interdit. Alors, que faire ? Le RFC 5892 (section 2.7) prévoyait une table des exceptions, où on pouvait mettre les caractères à qui on désirait épargner le sort commun. Fallait-il y mettre le chiffre 1 du Tai Lue ?

La décision finalement prise, et documentée par ce RFC (décision pompeusement baptisée « IETF consensus », bien que la section 5 rappelle qu'elle a été vigoureusement discutée) a finalement été de laisser s'accomplir le destin. Aucune exception n'a été ajoutée pour ces trois caractères, l'algorithme normal s'applique et U+19DA ne peut donc plus être utilisé. L'argument principal est que ce caractère n'était quasiment pas utilisé. Les cas suivants mèneront donc peut-être à des décisions différentes (la section 2 rappelle bien que, dans le futur, les choses seront peut-être différentes).

Ce RFC 6452 documente donc cette décision. Bien qu'elle se résume à « ne touchons à rien », il était préférable de la noter, car un changement qui casse la compatibilité ascendante n'est pas à prendre à la légère.

Les deux autres caractères, utilisés en Kannada, [Caractère Unicode non montré] (U+0CF1) et [Caractère Unicode non montré] (U+0CF2) ne créent pas le même problème car ils ont changé en sens inverse (interdits autrefois, ils sont désormais autorisés).