RFC 9233 : IDNA2008 and Unicode 12.0.0

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 24 avril 2022

Date de publication du RFC : Mars 2022

https://www.bortzmeyer.org/9233.html

La norme pour les noms de domaine en Unicode, dite « IDNA 2008 », prévoit une révision à chaque nouvelle version d'Unicode pour éventuellement s'adapter à des changements dus à ces nouvelles versions. Ce processus de révision a pas mal cafouillé (euphémisme), et ce RFC doit donc traiter d'un coup les versions 6 <https://www.bortzmeyer.org/unicode-6-0.html> à 12 <https://www.bortzmeyer.org/unicode-12-0.html> d'Unicode.

Le fond du problème est que l'ancienne norme sur les IDN (RFC 3490 [1]) était strictement liée à une version donnée d'Unicode et qu'il fallait donc une nouvelle norme pour chacune des versions annuelles d'Unicode. Vu le processus de publication d'une norme à l'IETF, ce n'était pas réaliste. La seconde norme IDN <https://www.bortzmeyer.org/idnabis.html>, « IDN 2 » ou « IDN 2008 » (bien qu'elle soit sortie en 2010) remplaçait les anciennes tables fixes de caractères autorisés ou interdits par un algorithme, à faire tourner à chaque sortie d'une version d'Unicode pour produire les tables listant les caractères qu'on peut accepter dans un nom de domaine internationalisé (le mécanisme exact, utilisant les propriétés des caractères listées dans la norme Unicode, figure dans le RFC 5892). En théorie, c'était très bien. En pratique, malgré les règles de stabilité d'Unicode, il se produisait parfois des problèmes. Comme le documente le RFC 8753, un caractère peut ainsi passer d'interdit à autorisé, ce qui n'est pas grave mais aussi dans certains cas d'autorisé à interdit ce qui est bien plus embêtant : que faire des noms déjà réservés qui utilisent ce caractère ? En général, il faut ajouter une exception manuelle, ce qui justifie un mécanisme de révision de la norme IDN, mis en place par ce RFC 8753. Ce nouveau RFC 9233 est le premier RFC de révision. Heureusement, cette fois, aucune exception manuelle n'a été nécessaire.

La précédente crise était due à Unicode version 6 <https://www.bortzmeyer.org/unicode-6-0.html> qui avait créé trois incompatibilités (RFC 6452). Une seule était vraiment gênante, le caractère

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc3490.txt

[Caractère Unicode non montré [2]] <https://r12a.github.io/uniview/?char=19DA>, qui, autorisé auparavant, était devenu interdit suite au changement de ses propriétés dans la norme Unicode. Le RFC 6452 avait documenté la décision de ne rien faire pour ce cas, ce caractère n'ayant apparemment jamais été utilisé. Unicode 7 <https://www.bortzmeyer.org/unicode-7-0.html> a ajouté un autre problème, le [Caractère Unicode non montré] <https://r12a.github.io/uniview/?char=08A1>, qui était un cas différent (la possibilité de l'écrire de plusieurs façons), et la décision a été prise de faire désormais un examen formel de chaque nouvelle version d'Unicode. Mais cet examen a été souvent retardé et voilà pourquoi, alors qu'Unicode 13 est sorti <https://www.bortzmeyer.org/unicode-13-0.html> (ainsi qu'Unicode 14 depuis <https://www.bortzmeyer.org/unicode-14-0.html>), ce nouveau RFC ne traite que jusqu'à la version 12 <https://www.bortzmeyer.org/unicode-12-0.html>.

Passons maintenant à l'examen des changements apportés par les versions 7 à 12 d'Unicode, fait en section 3 du RFC :

— À part le cas du [Caractère Unicode non montré] <https://r12a.github.io/uniview/?char=08A1> cité plus haut, Unicode 7 <https://www.bortzmeyer.org/unicode-7-0.html> n'a pas apporté de changements gênants pour IDN (par exemple, U+17B4 (caractère non visible) <https://r12a.github.io/uniview/?char=17B4> a changé de propriétés mais il était interdit pour IDN et le reste),
— Unicode 8 <https://www.bortzmeyer.org/unicode-8-0.html>, Unicode 9 <https://www.bortzmeyer.org/unicode-9-0.html> et Unicode 10 <https://www.bortzmeyer.org/unicode-10-0.html> n'ont appporté aucun changement gênant,
— Unicode 11 <https://www.bortzmeyer.org/unicode-11-0.html> a changé les propriétés de certains caractères existants mais le résultat pour IDN ne change pas (par exemple, [Caractère Unicode non montré] <https://r12a.github.io/uniview/?char=11A07>, qui était autorisé, le reste),
— Et Unicode 12 <https://www.bortzmeyer.org/unicode-12-0.html> ? Rien de problématique.

En Unicode 11, [Caractère Unicode non montré] <https://r12a.github.io/uniview/?char=111C9> qui passe d'interdit à autorisé, était un cas peu gênant. Le RFC prend donc la décision de ne pas ajouter d'exception pour ce caractère peu commun.

Voilà, arrivé ici, vous pensez peut-être que cela fait beaucoup de bruit pour rien puisque finalement les différentes versions d'Unicode n'ont pas créé de problème. Mais c'est justement pour s'assurer de cela que cet examen était nécessaire.

Pour compliquer davantage les choses, on notera qu'il existe encore sans doute (section 2.3 du RFC) des déploiements d'IDN qui en sont restés à la première version (celle du RFC 3490) voire qui sont un mélange des deux versions d'IDN, en ce qui concerne l'acceptation ou le refus des caractères.

En avril 2022, le travail pour Unicode 13 <https://www.bortzmeyer.org/unicode-13-0.html> ou Unicode 14 <https://www.bortzmeyer.org/unicode-14-0.html> n'a apparemment pas encore commencé...

2. Car trop difficile à faire afficher par LaTeX