RFC 5890 : Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 22 août 2010

Date de publication du RFC : Août 2010

https://www.bortzmeyer.org/5890.html

La nouvelle norme pour les noms de domaine écrits en Unicode, nommée IDNA2008, modifie les règles qui s'appliquent à ces IDN. Elle est composée de plusieurs RFC dont ce RFC 5890 [1], qui fixe la terminologie.

Dans cette nouvelle version, sur laquelle le travail a commencé en 2008 (d'où son nom officiel d'IDNA2008, cf. section 1.1), la norme est plus riche et plus complexe. Le seul RFC des définitions fait 29 pages. Elle remplace IDNA 1, ou « IDNA2003 » (l'ancienne norme, dans les RFC 3490 et RFC 3491). Elle crée notamment les nouveaux termes de "*U-label*" (composant de nom de domaine en Unicode) et de "*A-label*" (composant de nom de domaine internationalisé encodé selon l'algorithme Punycode du RFC 3492). Contrairement à d'autres normes de l'IETF, elle n'est pas indispensable que pour les programmeurs, mais aussi pour ceux qui, par exemple, décident des politiques d'enregistrement des registres.

Quels sont les RFC qui composent IDNA2008 ? La section 1.3 donne la liste officielle :
— Ce document, le RFC 5890, qui donne les définitions.
— Le RFC 5894, de justification des choix effectués, et qui fournit aussi des avis sur les politiques d'enregistrement. Il n'a pas statut de norme et reflète des opinions peu consensuelles sur les IDN.
— Le RFC 5891 qui normalise le protocole.
— Le RFC 5893, spécifique aux questions posées par les écritures de droite à gauche,
— Le futur RFC sur les fonctions qui transforment un nom de domaine avant de le passer à IDNA, par exemple pour mettre en œuvre les équivalences entre deux caractères.

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc5890.txt

— Et enfin la liste des caractères autorisés, dans le RFC 5892. C'est une des grandes nouveautés, puisque, contrairement à la précédente version, elle ne dépend plus d'une version particulière de la norme Unicode.

La section 2 est ensuite consacrée aux mots-clés de IDNA2008 (attention à ceux qui connaissaient l'ancienne norme, le vocabulaire est souvent nouveau). Les RFC de IDNA2008 utilisent également un vocabulaire non-spécifique aux IDN mais qui est rappelé en section 2.2. Ainsi, un **registre** désigne toute organisation qui enregistre des noms de domaine, même si elle ne gère pas un TLD (je suis le registre de `bortzmeyer.org`). LDH ("*Letters Digits Hyphen*") désigne les caractères ASCII qui sont traditionnellement autorisés dans les noms de machine (RFC 1123). Notez bien que les noms de domaine, contrairement à ce qu'écrivent beaucoup d'ignorants <`https://www.bortzmeyer.org/pourquoi-idn-et-pas-un-dns.html`>, ne sont **pas** limités à LDH.

La section 2.3 introduit les termes spécifiques à IDN. Par exemple :

— "*LDH label*" est le composant ("*label*") d'un nom de domaine, qui s'écrit uniquement avec LDH (c'est le nom traditionnel comme `example` dans `www.example.com`, nom qui compte trois composants). Deux sous-ensemble de "*LDH label*" sont définis, "*Reserved LDH label*" (ceux dont le troisième et quatrième caractères sont des tirets) et les non-réservés (les noms de domaines pré-IDN comme `bortzmeyer.org`). Parmi les réservés, certains ont `xn` en premier et deuxième caractère. Ils forment les "*XN labels*" dont tous, c'est important, ne sont pas forcément des encodages valides en Punycode. Ceux qui sont valides sont les "*A-labels*", les autre étant nommés d'un terme péjoratif absolument non justifié, "*fake A-labels*" (IDNA2008 contient beaucoup de réglements de compte via le vocabulaire). La figure 1 du RFC représente graphiquement les relations entre ces différents ensembles. Il est recommandé de la consulter en cas de migraine.

— Les "*A-labels*" sont donc la forme ASCII des IDN, produite par Punycode (RFC 3492). Par exemple, `xn--stphane-cya` est un "*A-label*".

— Un "*U-label*" est un composant valide d'un nom de domaine en Unicode. Par exemple, `stéphane` est un "*U-label*" (dont le "*A-label*" est le `xn--stphane-cya` cité plus haut). Toute chaîne Unicode n'est pas forcément un "*U-label*". Elle doit être normalisée en NFC et ne compter que des caractères autorisés. Tout "*U-label*" peut être converti en un "*A-label*" unique et réciproquement.

Notons que tous ces termes sont pour des composants d'un nom de domaine ("*label*"). Le nom lui-même, s'il contient au moins un "*A-label*" ou un "*U-label*" est un IDN.

Il y a plein d'autres détails sur les composants d'un nom. Par exemple, lorsque les normes IDNA2008 parlent de l'ordre d'un caractère, c'est une référence à l'ordre de transmission via le réseau. L'affichage peut être différent, puisque certaines écritures se font de droite à gauche.

Tout RFC doit comporter une section sur la sécurité et c'est ici la section 4. Avec IDN, il y a potentiellement des problèmes de débordement de tableau (le "*U-label*" peut avoir plus de caractères que son "*A-label*", section 4.2). Mais cette section est aussi l'occasion d'une attaque erronée contre les IDN, accusés d'augmenter la confusion des utilisateurs. D'où des conseils tout à fait inappropriés comme de montrer d'une manière spécifique les noms composés de plusieurs écritures (une pratique pourtant courante dans certains pays).

Les discussions sur cette section avaient été acharnées, avant même la création du groupe de travail, et ont donc mené à des paragraphes déroutants, bourrés d'allusions que le lecteur débutant ne comprendra sans doute pas (comme les mystérieux « risques », jamais explicités, de la section 4.4). Au moins, cette section et la 4.8 disent franchement que la question des caractères visuellement similaires n'a pas de solution technique.