RFC 6912 : Principles for Unicode Code Point Inclusion in Labels in the DNS

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 16 avril 2013

Date de publication du RFC : Avril 2013

https://www.bortzmeyer.org/6912.html

Ce RFC de l'IAB expose les principes que, selon l'IAB, devraient suivre les registres de noms de domaine lorsqu'ils décident quels caractères Unicode sont autorisés ou pas, pour l'enregistrement de noms de domaine internationalisés.

Notons tout de suite que les principes exposés dans ce RFC ne sont guère argumentés (les sections 6, 7 et 8, normalement consacrées à cette tâche, sont très pauvres). Dès l'introduction, le RFC dit qu'il ne faut pas autoriser tous les caractères mais ne dit pas pourquoi, ou quels dangers mystérieux menacent ceux qui oseraient passer outre. Mais, une fois qu'on a décidé de ne pas permettre tous les caractères Unicode (enfin, tous ceux qui sont légaux selon le RFC 5892 [1]), il reste le problème du choix. Il n'existe pas d'algorithme pour cela (le monde Unicode est bien trop complexe, reflétant la complexité des écritures du monde <https://www.bortzmeyer.org/worlds-writing-systems.html> et nos connaissances insuffisantes) et ce RFC, plutôt que de donner un tel algorithme, pose des principes qui devraient guider ceux et celles qui feront le choix.

La section 1 donne quelques objectifs (sans que la suite du RFC n'indique le rapport entre les principes et ces objectifs) : limiter les possibilités de confusion entre deux noms (par exemple google.com et goog1e.com, regardez bien le second), éviter qu'une adresse IP soit prise pour un nom de domaine comportant des éléments numériques (un faux problème typique, cf. RFC 1123, section 2.1), et favoriser l'accessibilité (cf. WCAG <http://www.w3.org/TR/WCAG/>). Le RFC introduit une notion importante, celle de **zone publique**. Une zone du DNS est un ensemble de sous-domaines contigus et gérés par la même organisation. Certaines sont privées, au sens où une seule organisation les utilisent (par exemple la zone univ-paris-diderot.fr n'est utilisée que par l'université Denis Diderot). D'autres sont publiques au sens où le registre qui la gère accepte des enregistrements de diverses organisations.

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc5892.txt

La plupart des TLD sont des zones publiques (par exemple `.fr`, `.org` ou `.pm`), ainsi que la racine. En effet, le RFC estime qu'il faut des règles particulières pour ces zones publiques (cf. section 4).

L'IAB s'est déjà exprimée sur cette question du choix des caractères autorisés pour les IDN. Notons que personne ne s'est posé la question pour les noms de domaines en ASCII alors que, comme le montre l'exemple `goog1e.com` plus haut, ils posent exactement les mêmes « problèmes ». Mais l'idée même de permettre à chacun, et pas seulement aux anglophones, d'écrire les noms de domaine avec leur écriture n'est toujours pas complètement acceptée. À part quelques ultra-réactionnaires, plus personne n'ose dire ouvertement qu'il est contre mais l'anti-IDNisme s'exprime plutôt aujourd'hui par un discours de peur, parlant de dangers (jamais clairement spécifiés) et laissant entendre qu'il faudrait contrôler sévèrement les IDN. Les déclarations précédentes de l'IAB peuvent être trouvées dans « *"IAB Statement : "The interpretation of rules in the ICANN gTLD Applicant Guidebook"* <http://www.iab.org/documents/correspondence-reports-documents/2012-2/iab-statement-the-interpretation-of-rules-in-> » et « *"Response to ICANN questions concerning 'The interpretation of rules in the ICANN gTLD Applicant Guidebook'"* <https://www.iab.org/documents/correspondence-reports-documents/2012-2/response-to-icann-questions-concerning-the-interpretation-of-rules-in-the-icann-gtld-> ». Dans ces textes, l'IAB plaidait pour un extrême conservatisme dans l'autorisation de caractères, en restreignant encore les règles techniques des RFC 5890 et RFC 5892, par exemple en n'autorisant que les caractères de la catégorie Unicode « Lettres ».

Dans ce RFC, l'IAB veut aller plus loin en interdisant également des caractères de cette catégorie. Le RFC prend l'exemple de [Caractère Unicode non montré [2]] (U+02BC, une apostrophe), qui a un rendu quasiment toujours identique à [Caractère Unicode non montré] (U+2019) alors qu'il est dans la catégorie Lettre (U+2019 étant dans la catégorie Ponctuation, mystères des catégories Unicode). Sans compter la traditionnelle ' (U+0027). Bien que, légalement, U+02BC soit un caractère autorisé dans un IDN, le RFC suggère que ce n'est sans doute pas une bonne idée de l'autoriser.

S'il y a des caractères dans la catégorie Lettres qui ne devraient sans doute pas être autorisés (cas ci-dessus), l'inverse existe aussi : des caractères situés dans d'autres catégories sont néanmoins indispensables dans certaines langues, notamment indiennes. Cela illustre le point mentionné plus haut : il n'y a pas d'algorithme pour établir automatiquement la liste des caractères autorisés, il va falloir y aller à la main. La section 4.2.4 du RFC 5891 mentionnait ce travail comme une reponsabilité indispensable du registre, et ce RFC 6912 ajoute que ce travail doit être fait à l'avance, pas décidé lorsqu'une demande d'enregistrement de nom se présente. Une telle décision lorsqu'une demande survient offrirait trop de possibilités à l'arbitraire. Non, il faut des règles pré-établies, et publiées. (À titre d'exemple, vous pouvez regarder les règles d'enregistrement des IDN dans .fr <http://www.afnic.fr/medias/documents/afnic-idn-specifications-techniques.pdf>.)

Le RFC estime que ces règles doivent être d'autant plus sévères (doivent autoriser moins de caractères) que l'on monte dans la hiérarchie des zones et que la zone racine doit donc avoir des règles particulièrement conservatrices, puisqu'elle concerne tous les utilisateurs de l'Internet.

Donc, pas d'algorithme (s'il était possible, il aurait sans doute déjà été développé) mais des principes. Quels sont-ils ? La section 3 en fournit une liste. D'abord, le principe de longévité : comme un caractère peut changer de catégorie, invalidant son usage (un exemple figure dans le RFC 6452), il est prudent de n'autoriser que des caractères qui sont stables depuis plusieurs versions d'Unicode.

Ensuite, le principe de moindre étonnement : un utilisateur normal ne devrait pas être trop surpris de la présence ou de l'absence de tel caractère. Notez que cela dépend du contexte : un utilisateur d'une écriture donnée ne sera pas surpris par les mêmes choses qu'un utilisateur d'une autre écriture.

2. Car trop difficile à faire afficher par LaTeX

Ces principes étaient valables pour toutes les zones. Mais les zones publiques ont des contraintes supplémentaires (section 4). Il y a le principe conservateur (dans le doute, rejeter le caractère), le principe d'inclusion (les caractères sont interdits par défaut et on inclut ensuite explicitement ceux qu'on estime « sûrs », terme que le RFC ne définit pas, et qui vient du FUD anti-Unicode), le principe de simplicité (un « honnête homme », avec des compétences modérées en DNS et en Unicode, doit pouvoir comprendre les raisons qui ont amené au choix d'acceptation ou de rejet, sans avoir eu besoin d'être présent aux quinze dernières réunions de l'IETF)...

Ce principe de simplicité dépend là encore du contexte. Si l'honnête homme cité plus haut ne connait aucune langue indienne, il ne comprendra sans doute pas les raisons qui ont mené à l'ajout de caractères non-lettres, indispensables pour certaines de ces langues. La racine servant à tous, ses règles devraient être ultra-simples et ne pas demander de compétences linguistiques particulières. Un ccTLD d'un pays donné peut se permettre des règles plus complexes pour des étrangers, mais qui sembleront simples à ceux qui connaissent les langues locales.

Pour les zones publiques, il y a aussi le principe de prédictabilité (les règles doivent donner un résultat identique, dès qu'on les applique sérieusement) et de stabilité (la liste des caractères autorisés ne devrait changer que rarement).

La racine, pour les raisons expliquées plus haut, a droit à une section spéciale (section 5). Le RFC demande qu'elle soit strictement peuplée de lettres, en s'appuyant sur une note de la section 2.1 du RFC 1123 qui dit que c'est le cas (c'était vrai lorsque le RFC 1123 a été publié, mais il n'y a aucun consensus à l'IETF sur l'interprétation de cette note : exprime t-elle une constatation ou une prescription ?) L'idée derrière cette restriction (RFC 4690) est que les noms de domaines n'ont pas vocation à permettre d'écrire tous les mots, encore moins des phrases correctes et complètes, mais uniquement celle de permettre la création de mnémoniques pratiques.

Ce RFC exprimant des opinions très contestables, la discussion avait été animée (voir par exemple les commentaires que j'avais fait <http://trac.tools.ietf.org/group/iab/trac/ticket/207> sur une précédente version, violemment anti-Unicode).