RFC 5992 : Internationalized Domain Names Registration and Administration Guideline for European languages using Cyrillic

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 5 octobre 2010

Date de publication du RFC : Octobre 2010

https://www.bortzmeyer.org/5992.html

Les noms de domaines internationalisés connaissent en ce moment un intérêt particulier, notamment en raison de leur déploiement (très tardif) dans la racine (ainsi, . [Caractère Unicode non montré [1]][Caractère Unicode non montré] a été ajouté le 12 mai 2010). Plusieurs RFC ont été publiés pour adapter la norme IDN ("*Internationalized Domain Names*") à différentes écritures et/ou langues et notre tout neuf RFC 5992 [2] s'ajoute à la liste, pour le cas de l'alphabet cyrillique.

Contrairement au RFC 5564 qui n'envisageait que le cas d'une seule langue, ce RFC 5992 veut couvrir la plupart des langues qui utilisent aujourd'hui une écriture, la cyrillique. Il s'applique donc au russe mais aussi au bulgare, à l'ukrainien, au serbe, etc.

Parmi toutes les écritures existantes, beaucoup ne servent que pour une seul langue. Ce n'est pas le cas de l'alphabet cyrillique, qui sert pour plusieurs langues slaves mais aussi pour des langues parlées dans l'ancien empire des tsars tel que le same de Kildin ou (bien que cela ne soit plus guère le cas depuis la chute de l'URSS) les langues « asiatiques » comme l'azéri ou le kirghiz. Ce RFC se focalise sur les langues utilisées en Europe, ce qui inclus les langues slaves et le cas particulier du same de Kildin.

À noter que le cyrillique, dérivé de l'alphabet grec, partage avec ce dernier et avec son cousin l'alphabet latin, plusieurs caractères, ce qui peut mener dans certains cas à des confusions. Les problèmes que pourraient poser celles-ci ne sont pas détaillés dans le RFC, qui fait sans doute allusion au FUD comme

1. Car trop difficile à faire afficher par LaTeX
2. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc5992.txt

quoi IDN augmenterait les risques de hameçonnage <https://www.bortzmeyer.org/idn-et-phishing.html>.

D'autre part, même dans les pays où l'écriture officielle est uniquement en cyrillique, et où son usage domine, on constate également une certaine présence de l'alphabet latin (par exemple pour les sigles techniques, regardez un article d'informatique écrit en russe, pour un exemple voir l'article du Wikipédia russophone sur le Web et ses mentions de HTML). Les utilisateurs souhaiteraient donc parfois pouvoir enregistrer des noms de domaine mêlant les deux alphabets, ce que le RFC déconseille fortement, sans expliquer pourquoi (section 1).

La section 1.1 expose la question des caractères « similaires » (terme impossible à définir rigoureusement, entre autres parce que cela dépend de la police d'affichage). Sans expliquer vraiment le problème, le RFC propose d'utiliser des mécanismes de « variantes » tels que présentés dans les RFC 3743 et RFC 4290.

Place maintenant aux langues et à leurs caractères. La section 2 décrit précisement les caractères nécessaires pour chaque langue. Vingt-trois « caractères de base » sont introduits, car ils sont communs à toutes les langues. Ils vont de U+0430 ([Caractère Unicode non montré]) à U+0448 (le cha, [Caractère Unicode non montré]), mais avec quelques trous. Ils incluent en outre les chiffres et le tiret. Il suffit donc ensuite de définir pour chaque langue les quelques caractères supplémentaires nécessaires. (On peut noter que la liste utilisée par le registre du `.eu` a 32 caractères <http://www.eurid.eu/en/eu-domain-names/technical-limitations/supported-characters>.)

Ainsi, le « serbo-bosnien » utilise les vingt-trois caractères de base plus sept autres, comme le U+045F (« dzhe », [Caractère Unicode non montré]). Le bulgare n'a pas ce caractère mais se sert entre autres du U+044F (ya, [Caractère Unicode non montré]). Le russe a besoin de trente-trois caractères en tout. Le RFC donne la liste complète pour chaque langue slave. Celles qui ne s'écrivent pas en cyrillique (ou qui ne s'écrivent **plus** dans cet alphabet, comme le moldave) ont été omises. À noter que, pour certaines langues, il existait déjà une liste officielle des caractères nécessaires. C'est le cas du monténégrin pour lequel la norme gouvernementale est <http://www.gov.me/files/1248442673.pdf>. En revanche, pour d'autres langues, il semble que l'orthographe ne soit pas complètement stabilisée. C'est le cas de l'ukrainien où les chiffres vont de trente-et-un à trente-quatre caractères (la dernière valeur étant recommandée par le RFC, pour être sûr de tout couvrir). Par exemple, certains affirment que le U+044A ([Caractère Unicode non montré]) est nécessaire, d'autres disent qu'il vaut mieux utiliser le U+044C ([Caractère Unicode non montré]) à la place. (À noter que les linguistes considèrent souvent que certaines langues n'en sont pas réellement, elles ont juste un nom différent pour des raisons politiques. Le monténégrin n'existe ainsi que depuis l'indépendance du Monténégro, tout le monde l'appelait serbe avant...)

Seule langue non-slave du groupe, le same de Kildin fait l'objet de la section 2.4. Son orthographe n'est pas normalisée et son système de sons est complexe, et difficile à rendre en cyrillique. Il faut donc rajouter trente-trois caractères au jeu de base comme par exemple le U+04CE ([Caractère Unicode non montré]). La section 2.4 contient des références vers des sources plus précises sur cette langue et son orthographe, si vous voulez approfondir la question.

Une fois ces listes établies, on peut compiler des tables qui indiquent les caractères autorisés (sections 3 et 4). La section 5 donne le format de la table qui figure dans l'annexe A du RFC : elle comprend au moins ligne de quatre colonnes par caractère. La première colonne indique le point de code Unicode du caractère cyrillique, la seconde son nom, la troisième et la quatrième les points de code et noms d'un caractère latin ou grec qu'on peut confondre avec le caractère cyrillique. Ainsi, U+0430 (le [Caractère Unicode non montré] cyrillique) est indiqué comme confondable avec le petit a latin, U+0061, mais aussi avec l'alpha grec, si on met les deux lettres en majuscules (le +++ dans la première colonne indique

que la confusion n'existe qu'en cas de changement de casse). De même, le U+0433 ([Caractère Unicode non montré], le ge) est indiqué comme proche du petit r, U+0072. Notez qu'il s'agit bien d'une confusion visuelle et pas sémantique. Le U+0440 ([Caractère Unicode non montré]) peut être confondu avec le p latin mais il représente un son complètement différent (proche du r).

Bien, armé de ces observations et de cette table, que peut faire un registre de noms de domaines qui accepte les caractères cyrilliques ? Prenons l'exemple classique du sigle russe de l'ancienne URSS, [Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré]. Ce sigle en cyrillique { U+0421 U+0421 U+0421 U+0420 } et le texte latin { U+0043 U+0043 U+0043 U+0050 } sont très proches. Alors, que doit faire un registre qui accepte les caractères cyrilliques et les latins (ce qui est le cas de `.eu`) ? Il peut allouer [Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré] et CCCP (le premier est en cyrillique, le second en latin) au même titulaire. Il peut n'autoriser qu'un des deux noms et bloquer l'autre. Ici, il n'y a que deux variantes. Mais, dans certains cas, il peut y en avoir plus (surtout en ajoutant l'alphabet grec ce qui, là encore, est le cas du `.eu`). Le registre peut alors choisir une solution intermédiaire en enregistrant certaines variantes et en bloquant d'autres. Ces possibilités, et leurs conséquences, sont discutées dans la section 6.