RFC 5893 : Right-to-left scripts for Internationalized Domain Names for Applications (IDNA)

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/5893.html

Le fait que certains systèmes d'écriture soient de gauche à droite (comme celui utilisé pour ce texte) et d'autres de droite à gauche ne pose pas de problèmes lorsque le texte est entièrement dans un sens ou dans l'autre. Mais, si on les mélange, on arrive parfois à des résultats surprenants <https://www.bortzmeyer.org/affichage-bidi.html>. Dans le contexte des noms de domaine, cela peut mener à rendre leur utilisation difficile. C'est pour cela que l'ancienne norme IDNA 1 limitait ce mélange (RFC 3491 [1], section 6, qui référence RFC 3454, section 6). Les limitations étaient un peu trop strictes et sont légèrement libéralisées par ce nouveau RFC 5893, qui fait partie de la nouvelle norme IDNAbis <https://www.bortzmeyer.org/idnabis.html>. Le changement est faible en pratique, la plupart des noms autorisés restent autorisés. En dépit d'une fréquente utilisation de "*weasel words*" <http://en.wikipedia.org/wiki/Wikipedia:Avoid_weasel_words#Unsupported_attributions> par ce RFC (comme « sûr » en section 1.3), il n'y a pas de conséquences, qu'elles soient positives ou négatives, pour la sécurité (malgré ce que raconte la section 9 du RFC).

On peut résumer l'ancienne norme (cf. section 1.2 de notre nouveau RFC) en disant que tout composant d'un nom de domaine ne devait **pas** inclure des caractères de directionnalité différente (par exemple de l'alphabet grec et de l'alphabet arabe) et qu'il devait commencer **et se terminer** par des caractères ayant une directionnalité déterminée (les chiffres arabes, par exemple, n'ont pas de directionnalité déterminée). Notons qu'il y a deux poids et deux mesures : les noms de domaine traditionnels en ASCII n'avaient pas cette limite et, par exemple, 7-septembre ou 3com sont des composants autorisés.

Il est prudent de relire la section 1.4 sur la terminologie, car tout le monde n'est pas expert BIDI. Soit on apprend par cœur le difficile UAX#9 <http://www.unicode.org/reports/tr9/>, la norme officielle du BIDI, soit on révise rapidement dans ce RFC les points importants :

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc3491.txt

— Les caractères Unicode ont tous une propriété BIDI : directionnalité de gauche à droite (les caractères de l'alphabet grec, par exemple), directionnalité de droite à gauche (les caractères de l'alphabet arabe), chiffres arabes (qu'Unicode appelle EN pour "*European Number*", comme 0, 1 ou 2, ils sont sans directionnalité puisqu'ils sont utilisés dans des alphabets des deux modèles), chiffres indo-arabes (qu'Unicode appelle AN pour "*Arabic Number*", comme [Caractère Unicode non montré [2]] (1) ou [Caractère Unicode non montré] (7) et qui ont une directionnalité dite « faible »), etc.
— Les chaînes de caractères ont deux ordres : l'ordre du réseau, qui est celui dans lequel les caractères de la chaîne ont été tapés, ou bien dans lequel ils sont transmis sur le réseau, et l'ordre d'affichage, qui est celui dans lequel ils sont présentés à des lecteurs humains. Lorsque le RFC parle d'ordre ou bien utilise des termes comme « premier » ou « dernier », c'est en général en référence à l'ordre du réseau,

Armé de ces définitions, on peut arriver au cœur du RFC, la section 2. Elle formalise les règles que doivent suivre des composants de noms de domaine internationalisés :

— Le composant d'un IDN doit commencer par un caractère de directionnalité forte (donc pas par un chiffre, cf. section 4.3 et 7.1) ; ce caractère détermine si le composant est gauche-à-droite ou droite-à-gauche,
— Dans un composant droite-à-gauche, seuls sont permis les caractères de directionnalité droite-à-gauche ou bien sans directionnalité (comme les chiffres). Ainsi, `[Caractère Unicode non montré ][Caractère Unicode non montré ][Caractère Unicode non montré ][Caractè Unicode non montré ]-XML-[Caractère Unicode non montré ][Caractère Unicode non montré ][Caractère Unicode non montré ][Caractère Unicode non montré ][Caractère Unicode non montré ]` (tiré de la documentation en arabe de SPIP) serait interdit, à cause du sigle en caractères latins (même si la présence de tels sigles est très courante dans les textes techniques en arabe),
— Le caractère final peut être sans directionnalité (on peut finir par un chiffre),
— Le mélange des chiffres arabes et indo-arabes dans un même label est interdit (notons que cette règle était déjà dans le RFC 5564) : les chiffres indo-arabes sont interdits dans un composant gauche-à-droite,
— Et je passe quelques règles plus subtiles.

Ce RFC 5893 propose également des justifications pour le choix de ces règles, sous forme de critères que devraient respecter tous les noms de domaine en Unicode (section 3) :

— Unicité du composant : deux composants distincts, affichés dans le même paragraphe, ne doivent **pas** avoir le même affichage. Sans les règles plus haut, les composants `123` et `321`, par exemple, pourraient s'afficher de manière identique, si le second est précédé de caractères droite-à-gauche. L'interdiction des chiffres (caractères sans directionnalité) au début d'un composant découle de ce critère.
— Regroupement des caractères : les caractères d'un même composant doivent rester groupés, ce qui ne serait pas le cas si on permettait le mélange de caractères de directionnalité différentes.

Dans le cours de la discussion sur IDNAbis, d'autres critères avaient été suggérés mais n'ont finalement pas été retenus :

— Constance de l'apparence : un composant doit être affiché de manière identique dans un contexte de gauche-à-droite et dans un contexte de droite-à-gauche : c'est un résultat trop difficile à obtenir dans le contexte de l'algorithme BIDI,
— L'ordre des composants d'un nom doit rester identique quel que soit le contexte d'affichage ; ce critère aurait mené à des tests inter-composants, peu réalistes (puisque ce sont des registres différents qui sont impliqués pour chaque niveau du nom, avec des règles différentes),
— Unicité du nom de domaine : deux noms différents ne devraient pas être affichés de manière identique ; objectif impossible à atteindre : `ABC.abc` sera affiché `abc.CBA` dans un contexte droite-à-gauche, et le nom différent `abc.ABC` sera affiché de manière identique dans un contexte gauche-à-droite.

2. Car trop difficile à faire afficher par LaTeX

Arrivé là, on a toutes les règles (la fin de la section 3 les reformalise de manière plus rigoureuse). La section 4 donne simplement des exemples de cas où les règles des RFC 3454 et RFC 3491 donnaient des résultats peu satisfaisants. Ainsi, la langue divehi, qui s'écrit avec un alphabet proche de l'arabe, le Thaana, a tous ses mots qui se terminent par un caractère Unicode combinant (un accent, disons en simplifiant). « Ordinateur » se dit en dhivehi « "*[Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré]*" » et le dernier caractère, U+07AA est l'"*ubu fili*", un caractère (pas une lettre) sans directionnalité, qui aurait été rejeté par IDNA 1 (section 4.1).

Un problème analogue se pose en yiddish. Ainsi, l'organisation qui normalise les règles d'écriture du yiddish s'écrit « "*[Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré]*" » et le dernier caractère, U+05B8, n'est pas une lettre (section 4.2).

Il n'existe pas de solution technique aux problèmes d'affichage BIDI, l'ensemble des situations possibles étant trop vaste. Il ne faut donc pas croire qu'appliquer les règles de ce RFC suffira à être tranquille. La section 5 donne quelques exemples, par exemple un nom de plusieurs composants, où un composant un IDN (satisfaisant les règles de ce RFC), précède des noms ASCII commençant par un chiffre : `[Caractère Unicode non montré ][Caractère Unicode non montré ][Caractère Unicode non montré ][Caractère Unicode non montré ][Caractère Unicode non montré ][Caractère Unicode non montré ][Caractère Unicode non montré ][Caractère Unicode non montré ].3com.com` va ainsi être affiché d'une manière déroutante (cela devrait être `[Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré].3com.com`). Ce nom n'est pas interdit (alors que c'était l'ambition initiale du groupe de travail idnabis <`http://tools.ietf.org/wg/idnabis`>) car il existe déjà beaucoup de noms ASCII commençant par un chiffre, et car la combinaison de composants pour former un nom est parfois réalisée automatiquement (par exemple via la directive `search` dans `/etc/resolv.conf`), ne laissant pas de possibilité de contrôle, et enfin parce que les jokers du DNS (encore eux) peuvent faire qu'un nom peut être résolu sans avoir été enregistré (et donc vérifié)...

La section 6 liste d'ailleurs quelques autres problèmes comme le fait que le mélange de chiffres arabes et de chiffres indo-arabes est interdit, mais que le mélange de chiffres bengalis et chiffres gujaratis n'est pas mentionné... Le cas doit être traité par le registre (par exemple celui de `.IN`).

Les règles de ce RFC étant nouvelles, il y a potentiellement des problèmes avec les anciens noms. La section 7.1 analyse les questions de compatibilité. La 7.2 se préoccupe au contraire du futur en constatant que les propriétés BIDI ne font **pas** partie des propriétés qu'Unicode s'engage à ne pas modifier et que donc, dans le futur, un changement de propriétés BIDI pourrait rendre invalides des composants valides et réciproquement.

Les IDN BIDI posent-ils des problèmes de sécurité particuliers ? C'est ce que laisse entendre Patrik F[Caractère Unicode non montré]ltstr[Caractère Unicode non montré]m dans son article « "*Mixing different scripts is hard*" <`http://stupid.domain.name/node/681`> », qui est franchement tendancieux. Si les exemples donnés d'affichage BIDI suprenants sont amusants intellectuellement, il n'est jamais démontré que cela puisse avoir des conséquences de sécurité. La section 9 de ce RFC 5893, consacrée à ce sujet, ne fournit pas d'éléments nouveaux, à part de vagues accusations.