

RFC 6927 : Variants in Second-Level Names Registered in Top Level Domains

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 7 mai 2013

Date de publication du RFC : Mai 2013

<https://www.bortzmeyer.org/6927.html>

Beaucoup d'utilisateurs des noms de domaine pensent que deux noms qui, pour eux, sont « le même », devraient être traitées d'un bloc par le DNS et/ou par les systèmes d'enregistrement de noms. Tout le problème est évidemment de définir « le même » : il n'y a en général pas d'accord entre les utilisateurs sur ce point. Ainsi, des anglophones peuvent penser que `meter.example` et `metre.example` sont « le même nom » et devraient donc se comporter ainsi. D'autres pensent que `pizzeria-napoli.example` et `pizzerianapoli.example` sont « le même nom ». Et, évidemment, depuis l'arrivée des IDN, le potentiel pour des noms « identiques » a encore augmenté. Il a donc été proposé que ces **variantes** d'un « même » nom soient explicitement reconnues. Ce RFC ne suggère pas de mécanisme en ce genre (je pense personnellement que le problème est bien trop flou et subjectif pour cela). Il se contente de regarder les pratiques actuellement déployées par les registres en matière de « variantes ».

Des exemples de caractères qui sont une « variante » d'un autre ne manquent pas dans les écritures du monde. Ainsi, la Chine a connu sous Mao une réforme orthographique, passant des sinogrammes « traditionnels » aux « simplifiés ». La plupart des sinophones considèrent que [Caractère Unicode non montré ¹] et [Caractère Unicode non montré] sont des variantes du même mot. Dans les langues utilisant l'alphabet latin avec des accents, certains considèrent que le mot avec et sans accent sont des variantes l'un de l'autre. Le RFC note qu'en français certains omettent les accents sur les majuscules (c'est incorrect, mais cela se fait parfois) et que donc `INTERNÉ` et `INTERNE` sont considérés comme « équivalents ». Ce qui est sûr, c'est qu'historiquement, les difficultés à manier les caractères non-ASCII dans beaucoup d'environnements informatiques avaient mené bien des utilisateurs à considérer les accents comme purement décoratifs, et pouvant être omis.

Est-ce que ce vague sentiment que deux groupes de caractères sont « la même chose » peut être formalisé dans une définition rigoureuse des « variantes » ? Il y a eu quelques tentatives comme à

1. Car trop difficile à faire afficher par L^AT_EX

l'ICANN <<https://www.icann.org/en/resources/idn/variant-tlds>>, ou à l'IETF dans le RFC 3743² (non standard). Mais un autre document <<https://www.icann.org/en/topics/idn/idn-vip-integrated-issues-final-clean-20feb12-en.pdf>> de la même ICANN concluait qu'il n'existait pas de définition utilisable. Ensuite, sur le comportement attendu de deux noms identiques ne fait pas non plus l'objet d'un consensus. Par exemple, si on veut, non pas seulement une réponse identique dans le DNS, mais un vécu de l'utilisateur identique, par exemple sur le Web, il n'y a aucune solution purement DNS (en raison du comportement de HTTP avec le champ `Host` :). Si vous voulez lire davantage sur cette question, je vous renvoie à mon article « Que veut dire synchroniser deux domaines? » <<https://www.bortzmeyer.org/domaines-synchrones.html>> ».

Ce RFC n'essaie donc pas de régler ce problème (très ancien et sans doute insoluble). Il fait un tour d'horizon des pratiques des gTLD (qui sont contractuellement obligés par l'ICANN de rédiger leurs pratiques IDN et de les déposer à l'ICANN <<https://www.iana.org/domains/idn-tables>>, et qui rendent accessibles le contenu de leurs zones, permettant aux chercheurs de l'étudier) et d'un certain nombre de ccTLD. Ces obligations de l'ICANN sont décrites dans « *Guidelines for the Implementation of Internationalized Domain Names* » <<https://www.icann.org/en/resources/idn/idn-guidelines-02sep11-en.htm>> ». Le RFC note que ces règles ne sont guère respectées, les politiques déposées à l'ICANN étant souvent dépassées et plus à jour (URL cassés, personnes désormais parties, etc). (Mon opinion personnelle est que ces « *IDN Guidelines* » n'ont aucun sens et sont motivées par des frayeurs bureaucratiques plutôt que par des considérations pratiques.)

Pensez à réviser la terminologie en section 2. Un **lot** ("*bundle*", on dit aussi "*IDL package*") est un ensemble de noms qui sont tous des variantes les uns des autres (RFC 3743). Un nom est **alloué** ("*allocated*") ou enregistré s'il est attribué à un titulaire. Il est **actif** ("*active*") s'il apparaît dans le DNS. Il est **bloqué** ("*blocked*") s'il ne peut pas être alloué du tout et **réservé** ("*withheld*") s'il ne peut être alloué que dans certaines conditions.

La section 4 décrit les politiques des gTLD (avec vérification dans les fichiers de zone de la réalité de ces politiques). Certains sont faciles : `.aero`, `.coop` ou `.pro` n'ont pas d'IDN (même chose pour `.xxx`). D'autres ont un grand intérêt pour les IDN. Ainsi, `.asia` a évidemment une politique IDN détaillée <<http://dot.asia/policies/DotAsia-CJK-IDN-Policies-COMLETE--2011-05-04.pdf>>, avec gestion de lots, via les enregistrements NS (les variantes ont le même jeu de NS). `.cat` a un cas plus simple que `.asia` puisqu'il ne reconnaît qu'une seule langue, le catalan. La liste des caractères autorisés est publiée <<https://www.icann.org/en/about/agreements/registries/cat/cat-appendixs-22mar08.htm>> et elle comprend des caractères comme le l·l (considéré comme un caractère en catalan mais représenté par trois caractères Unicode). Lorsqu'on enregistre un nom IDN, l'ASCII « équivalent » est aussi enregistré et un DNAME (RFC 6672) est mis dans le fichier de zones (il semble que cela soit le seul gTLD à procéder ainsi). Par exemple :

```
xn--caball-gva.cat. 43200 IN DNAME caballe.cat.
```

Le nom IDN `caballé.cat` pointant vers l'ASCII `caballe.cat` (et merci à DNSDB <<https://www.bortzmeyer.org/dnsdb.html>> pour permettre l'exploration facile du DNS à la recherche d'exemples).

Contrairement à `.asia` et a fortiori à `.cat`, `.com` doit gérer toutes les écritures du monde. VeriSign a transmis à l'ICANN des tables pour des écritures hiéroglyphiques et pour l'araméen... Les politiques suivies sont décrites en détail dans plusieurs documents (voir la liste dans la bibliographie du RFC).

2. Pour voir le RFC de numéro NNN, <https://www.ietf.org/rfc/rfcNNN.txt>, par exemple <https://www.ietf.org/rfc/rfc3743.txt>

Pour l'écriture chinoise, et pour elle seule, les variantes d'un nom alloué sont bloquées. (.net et .name suivent la même politique que .com.)

.org publie une liste des langues gérées <<http://www.pir.org/why/global>> (notez que c'est absurde puisque IDN gère des écritures, pas des langues, mais cette confusion est fréquente dans les cercles ICANN, où il est toujours mal admis que des gens osent utiliser d'autres écritures que celle des États-Unis). Les tables déposées à l'ICANN sont un sous-ensemble de ces tables publiées par le registre (pas de mise à jour?) Les documentations officielles de .org font allusion à une politique de gestion des variantes mais apparemment sans dire en quoi consiste cette politique.

.tel a beaucoup d'IDN et une politique publique <<http://www.telnic.org/policies.html>>. Les noms avec l'écriture japonaise utilisent un système de blocage (le premier arrivé empêche l'allocation des variantes) alors que pour l'écriture chinoise, les variantes sont publiées (avec des enregistrements NS identiques).

La section 5 passe aux ccTLD. Je ne vais pas tous les citer mais ils ont souvent des politiques intéressantes. .cl publie sa politique <<http://www.nic.cl/CL-IDN-policy.html>>, qui ne tient pas compte d'éventuelles variantes.

.cn a évidemment plus de travail : leur politique est publiée sous forme d'un RFC, le RFC 4713. Les variantes sont publiées, avec des enregistrements NS. Par contre, .ir n'a pas d'IDN <http://www.nic.ir/Internationalized_Domain_Names>. Comme un certain nombre de registres des pays utilisant l'écriture arabe (qui va de droite à gauche), ils considèrent que le mélange d'un nom en écriture arabe avec un TLD en écriture latine (allant de gauche à droite) n'est pas souhaitable. (Mais il y a des IDN <<http://www.nic.ir/IDN#Dot-Iran>> dans le TLD équivalent en écriture arabe, [Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré].)

Bien que TLD d'un pays majoritairement anglophone, .nz a une politique IDN <<http://nzrs.net.nz/dns/idn>> (qui ne parle pas des variantes) pour le maori. On notera que le domaine compte des DNAME par exemple m[Caractère Unicode non montré]ori.dns.net.nz qui pointe vers l'équivalent ASCII :

```
xn--mori-gsa.dns.net.nz. 86400 IN DNAME maori.dns.net.nz.
```

Contrairement aux autres ccTLD, il semble que .ru ne publie pas sa politique directement, uniquement via la table déposée à l'ICANN.

La politique utilisée par .fr n'est pas étudiée dans ce RFC. Elle est disponible en ligne <<http://www.afnic.fr/medias/documents/afnic-idn-specifications-techniques.pdf>>. La liste des caractères acceptés figure en section 2 (ce sont les caractères utilisés pour les langues d'Europe de l'Ouest). S'il y a eu un système de variantes pendant la période d'ouverture (section 4), il a ensuite disparu.

Voilà, si vous voulez approfondir la question et voir la politique des autres TLD, consulter ce RFC. J'espère en tout cas avoir donné une bonne idée de la variété des écritures du monde et des solutions pour les gérer dans les noms de domaine.