RFC 8228 : Guidance on Designing Label Generation Rulesets (LGRs) Supporting Variant Labels

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 23 août 2017

Date de publication du RFC : Août 2017

https://www.bortzmeyer.org/8228.html

Le RFC 7940 [1] définissait un langage, LGR ("*Label Generation Ruleset*"), pour exprimer des politiques d'enregistrement de noms écrits en Unicode, avec les éventuelles **variantes**. Par exemple, avec LGR, on peut exprimer une politique disant que les caractères de l'alphabet latin et eux seuls sont autorisés, et que « é » est considéré équivalent à « e », « ç » à « c », etc. Ce nouveau RFC donne quelques conseils aux auteurs LGR. (Personnellement, je n'ai pas trouvé qu'il apportait grand'chose par rapport au RFC 7940.)

LGR peut s'appliquer à tous les cas où on a des identificateurs en Unicode mais il a surtout été développé pour les IDN (RFC 5890). Les exemples de ce RFC 8228 sont donc surtout pris dans ce domaine d'application. Un ensemble de règles exprimées dans le langage du RFC 7940 va définir deux choses :

— Si un nom donné est acceptable ou pas (dans l'exemple donné plus haut, un registre de noms de domaine ouest-européen peut décider de n'autoriser que l'alphabet latin),
— Et quels noms sont des **variantes** les uns des autres. Une variante est un nom composé de caractères Unicode différents mais qui est considéré comme « le même » (la définition de ce que c'est que « le même » n'étant pas technique ; par exemple, un anglophone peut considérer que "*theatre*" et "*theater*" sont « le même » mot, alors que, du point de vue technique, ce sont deux chaînes de caractère différentes). Le RFC dit que deux noms sont « le même » s'ils sont visuellement similaires (`google` vs. `google`), mais c'est une notion bien trop floue, et qui dépend en outre de la police de caractères utilisée (avec Courier, `google` et `google` sont difficilement distinguables). Le RFC ajoute qu'il y a également le cas de la similarité sémantique (le cas de "*theatre*" vs. "*theater*", ou bien celui des sinogrammes traditionnels vs. les simplifiés ou encore, mais c'est compliqué, le cas des accents en français, avec « café » et « cafe » mais aussi avec « interne » et « interné »).

1. Pour voir le RFC de numéro NNN, `https://www.ietf.org/rfc/rfcNNN.txt`, par exemple `https://www.ietf.org/rfc/rfc7940.txt`

L'idée est que le registre utilise ensuite ces règles pour accepter ou refuser un nom et, si le nom est accepté, pour définir quelles variantes deviennent ainsi bloquées, ou bien attribuées automatiquement au même titulaire. **Notez une nouvelle fois qu'il s'agit d'une décision politique et non technique.** Le RFC dit d'ailleurs prudemment dans la section 1 qu'il ne sait pas si ce concept de variante est vraiment utile à quoi que ce soit. (J'ajoute qu'en général ce soi-disant risque de confusion a surtout été utilisé comme FUD anti-Unicode, par des gens qui n'ont jamais digéré la variété des écritures humaines.)

J'ai dit qu'une variante était « le même » que le nom de base. Du point de vue mathématique, la relation « est une variante de » doit donc être symétrique et transitive (section 2). « Est une variante de » est noté avec un tiret donc `A - B` signifie « A est une variante de B » (et réciproquement, vu la symétrie). Notez que le RFC 7940 utilisait un langage fondé sur XML alors que ce RFC 8228 préfère une notation plus compacte. La section 18 de notre RFC indique la correspondance entre la notation du RFC 7940 et celle du nouveau RFC.

Mais toutes les relations entre noms « proches » ne sont pas forcément symétriques et transitives. Ainsi, la relation « ressemble à » (le chiffre 1 ressemble à la lettre l minuscule) est non seulement vague (cela dépend en effet de la police) mais également non transitive (deux noms ressemblent chacun à un nom dont la forme est quelque part entre eux, sans pour autant se ressembler entre eux).

Le RFC 7940 permet de définir des relations qui ne sont ni symétriques, ni transitives. Mais ce nouveau RFC 8228 préfère les relations d'équivalence.

Bon, et maintenant, comment on crée les variantes d'un nom ou d'un composant d'un nom (section 6) ? On remplace chaque caractère du nom originel par toutes ses variantes possibles (le fait qu'un caractère correspond à un autre est noté –¿). Ainsi, si on a décidé que « e » et « é » étaient équivalents (é –¿ e), le nom « interne » aura comme variantes possibles « interné », « intérne » et « intérné » (oui, une seule de ces variantes a un sens en français, je sais). Une fois cette génération de variantes faites, le registre, selon sa politique, pourra l'utiliser pour, par exemple, refuser l'enregistrement d'« interné » si « interne » est déjà pris (je le répète parce que c'est important : ce RFC ne décrit pas de politique, l'exemple ci-dessus n'est qu'un exemple).

En pratique, travailler caractère par caractère n'est pas toujours réaliste. Il y a des cas où c'est un groupe de caractères qui poserait problème. La section 7 introduit une amélioration où on peut étiqueter les correspondances de manière asymétrique. Le caractère x veut dire que l'allocation du nom est interdite (détails en section 9), le a qu'elle est autorisée (détails en section 8). On pourrait donc avoir « e x–¿ é » et « é a–¿ e » ce qui veut dire que la réservation de « interne » bloquerait celle d'« interné » mais pas le contraire.

Le monde des identificateurs étant très riche et complexe, on ne s'étonnera pas que des règles trop simples soient souvent prises en défaut. Ainsi, la variante peut dépendre du contexte : dans certaines écritures (comme l'arabe), la forme d'un caractère dépend de sa position dans le mot, et un même caractère peut donc entrainer de la confusion s'il est à la fin d'un mot mais pas s'il est au début. Il faut donc étiqueter les relations avec ce détail (par exemple, « final : C –¿ D » si le caractère noté C peut se confondre avec le caractère D mais uniquement si C est en fin de mot, cf. section 15 du RFC).

Si vous développez du LGR, au moins deux logiciels peuvent aider, celui de Viagénie <http://www.viagenie.com/>, lgr-crore <https://github.com/icann/lgr-core> et lgr-django <https://github.com/icann/lgr-django> et un logiciel de l'auteur de ce RFC, développé pour le « "*ICANN Integration Panel work*" » mais qui ne semble pas publié.