RFC 6067 : BCP 47 Extension U

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 8 décembre 2010

Date de publication du RFC : Décembre 2010

https://www.bortzmeyer.org/6067.html

Le RFC 5646 [1] (alias « BCP 47 » pour "*Best Common Practice 47*"), qui normalise les étiquettes de langue <http://www.langtag.net/>, prévoyait un mécanisme d'extension par le biais de sous-étiquettes d'un seul caractère. Ce RFC 6067 spécifie la première de ces extensions à rentrer en service, pour indiquer l'information de localisation du consortium Unicode.

Les étiquettes de langue sont utilisées pour marquer du contenu sur le Web mais également en plein d'autres endroits, afin d'indiquer la langue utilisée. Leur norme, le RFC 5646 décrit un mini-langage permettant d'indiquer la langue mais aussi l'écriture, le pays, voire la variante dialectale. Ainsi, `ru-petr1708` désignera le russe tel qu'il était écrit dans l'ortographe de Pierre Ier, avant la réforme de 1917. Ce langage de construction des étiquettes est très riche mais ne permet de faire une étiquette qu'à partir de sous-étiquettes déjà enregistrées dans le registre des langues <http://www.langtag.net/registries.html> (ou bien à partir de sous-étiquettes purement privées). Il n'y a notamment pas de moyen pour utiliser les catalogues existants.

Or, un de ces catalogues est particulièrement utilisé, le catalogue des "*locales*" décrit dans le TR35 <http://unicode.org/reports/tr35/> et géré par le consortium Unicode <http://www.unicode.org/>. C'est pour pouvoir l'utiliser que l'extension « u » est créée par notre RFC. (Le singleton « u » voulant dire Unicode.) Il permettra d'étiqueter avec davantage de précision un document. Ainsi, `en-u-cu-usd` désignera un texte en anglais dont, grâce à l'extension « u », on pourra savoir qu'il utilise le dollar états-unien (`usd`) comme unité monétaire.

Les données utilisables avec cette extension proviennent du CLDR, le grand dépôt des "*locales*" géré par le consortium Unicode et qui contient des choses aussi variées que les jours fériés par pays ou bien les différents ordres de tri utilisés.

1. Pour voir le RFC de numéro NNN, `https://www.ietf.org/rfc/rfcNNN.txt`, par exemple `https://www.ietf.org/rfc/rfc5646.txt`

Le RFC 5646/BCP 47 (section 3.7) impose un certain nombre de règles pour la création d'une extension, notamment l'indication précise de l'autorité en charge du catalogue accessible via l'extension, et de ses politiques. La section 2 de notre RFC 6067 satisfait à cette règle en décrivant comment est géré CLDR.

Maintenant, quelle information peut-on indiquer avec l'extension « u »? La section 2.1 les liste en renvoyant à la section 3 du TR35. On peut indiquer des attributs, des clés et des types. Aujourd'hui, aucun attribut n'est défini. Les clés, elles, ont exactement deux caractères et sont définis par le TR35. `ca` désigne un calendrier, `co` un ordre de tri, `cu` la monnaie, `tz` le fuseau horaire, etc. Les types sont les valeurs associées aux clés. Ainsi, `ca-coptic` désigne le calendrier copte. Une étiquette complète comme `de-DE-u-co-phonebk` sera « l'allemand tel qu'écrit en Allemagne, utilisant l'ordre de tri `phonebk`, i.e. celui normalisé pour l'annuaire téléphonique (qui se nomme `phonebook` dans CLDR, qui n'a pas les mêmes contraintes de taille) ». `en-u-tz-usden` sera l'anglais avec le fuseau horaire " *Mountain Time*". Et `es-u-cu-mxn` sera l'espagnol avec comme unité monétaire le peso mexicain. Bien sûr, dans la plupart des cas, il n'y aura pas besoin d'étiqueter les textes avec ce niveau de précision. (Merci à Doug Ewell pour la plupart des exemples.) Mais certaines utilisations pourront en avoir besoin.

CLDR distribue des fichiers contenant les informations nécessaires pour tous les types possibles en `<http://unicode.org/Public/cldr/>`. Si vous voulez l'ordre de tri allemand, il est en `common/collation/de`

La section 2.2 du RFC contient le formulaire d'enregistrement obligatoire (RFC 5646, section 3.7) pour une extension. « u » est donc désormais le premier élément du registre des extensions. `<https://www.iana.org/assignments/language-tag-extensions-registry>`.

Attention à un petit piège : les extensions comme « u » n'ont rien à voir avec les "*Extended Language Subtags*" (alias extlangs), qui sont un mécanisme (pas le seul `<https://www.bortzmeyer.org/extlang-or-not-extlang.html>`) pour représenter des idiomes intermédiaires entre une « vraie » langue et un dialecte.