RFC 5645 : Update to the Language Subtag Registry

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 7 septembre 2009

Date de publication du RFC : Septembre 2009

https://www.bortzmeyer.org/5645.html

Ce RFC, compagnon du RFC 5646 [1], décrit le nouvel état du registre des langues de l'IANA. Comme son prédécesseur, le RFC 4645 avait servi à initialiser le registre des langues, notre RFC 5645 servira à la grande mise à jour qu'est l'intégration des normes ISO-639-3 <https://www.bortzmeyer.org/iso-639-3.html> et ISO 639-5 <https://www.bortzmeyer.org/iso-639-5.html>. Le nombre de langues enregistrées passe de 500 à plus de 7 000.

Le registre est écrit dans le format dit "*record-jar*", décrit dans le livre "*The Art of Unix programming*" <https://www.bortzmeyer.org/art-unix-programming.html>.

Il sert juste à lister l'état du nouveau registre IANA <https://www.iana.org/assignments/language-subtag-registry> avec des affectations comme ici pour le Nandi ou l'écriture du Bamum :

```
%% Description d'une langue, ici le Nandi (niq)
Type: language
Subtag: niq
Description: Nandi
Added: 2009-07-29
Macrolanguage: kln
%%
%% Description d'une écriture, ici le Bamum
Type: script
Subtag: Bamu
Description: Bamum
Added: 2009-07-30
%%
```

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 registre va ensuite vivre sa vie et accepter de nouveaux enregistrements, suivant les procédures du RFC 5646. Une des façons de suivre ces futures évolution est de s'abonner au flux de syndication <http://www.langtag.net/registries/lsr.atom>.

Comment a été construit le nouveau registre ? La section 2 répond à cette question. Le point de départ était le registre existant, celui qui avait été créé par le RFC 4646. Les fichiers d'ISO 639-3 (disponibles <http://www.sil.org/iso639-3/download.asp> chez SIL car l'ISO ne distribue quasiment jamais ses normes) et ISO 639-5 ont ensuite été intégrés, mais pas aveuglément :

— Les langues qui n'avaient pas déjà de code dans le registre (comme l'ankave - code aak - ou le ghotuo - code aaa) ont été ajoutées. "*A contrario*", celles qui avaient déjà un code ont été ignorées.
— Les langues qui avaient une **macrolangue** (une nouveauté d'ISO 639-3, la macrolangue est une catégorie regroupant des langues qui sont parfois considérées comme distinctes et parfois comme une seule langue) ont été traitées comme les autres à l'exception de six langues pour lesquelles le groupe de travail LTRU de l'IETF a estimé qu'elles présentaient des caractéristiques particulières, notamment le fait que la macrolangue était souvent utilisée. Ces six exceptions sont l'arabe (ar), le konkani (kok), le malais (ms), le swahili (sw), l'ouzbèque (uz) et le chinois (zh). Pour les six exceptions, un "*extended language subtag*" a été créé dans le registre, pour indiquer la macrolangue et cela permettra d'écrire des étiquettes comme zh-yue (le cantonais, dont l'étiquette canonique est juste yue) pour le cas où l'ajout de la macrolangue semble important. Les autres langues ayant une macrolangue voient juste cette macrolangue ajoutée comme un champ de l'enregistrement.
— Les descriptions présentes dans le registre ont parfois été complétées ou modifiées pour avoir une présentation cohérente. Cela aboutit parfois à élargir la portée d'un code. Ainsi, la collection afa « Autres langues afro-asiatiques » est devenue « Langues afro-asiatiques » ce qui est bien plus large.
— Le RFC 5646 a inclus dans les codes de pays les éléments de ISO 3166 qui étaient marqués comme « exceptionellement réservés ». Cela a permis l'arrivée de EU, on peut donc désormais utiliser l'étiquette en-EU pour l'anglais parlé à Bruxelles. D'autres codes étaient déjà présents pour d'autres raisons comme l'amusant FX (France Métropolitaine).