RFC 5646 : Tags for Identifying Languages

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/5646.html

Successeur du RFC 4646[1] (qui lui-même succédait au RFC 3066), notre RFC décrit les noms des langues (langues humaines, pas langages informatiques). Toute application qui a besoin d'indiquer une langue doit s'en servir.

Le protocole HTTP par exemple permet à un navigateur Web d'indiquer au serveur quelle langue il préfère (en-tête `Accept-Language:` dans le RFC 7231, section 5.3.5), au cas où le serveur aie plusieurs versions d'une page et soit correctement configuré (ce que Apache permet <http://httpd.apache.org/docs/content-negotiation.html>). Cela marche très bien avec des sites comme <http://www.debian.org/>.

Des normes non-IETF (notamment XML) se réfèrent à ce RFC.

Les noms des langues ont deux ou trois lettres et sont tirés de la norme ISO 639. Par exemple :

— "fr" pour le français (à ne pas confondre avec le TLD ".fr", qui désigne la France),
— "ar" pour l'arabe,
— "br" pour le breton,
— "en" pour l'anglais,
— etc.

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc4646.txt

La norme complète est plus complexe : par exemple, l'anglais parlé aux États-Unis n'est pas tout à fait le même que celui parlé en Grande-Bretagne. Aussi, notre RFC permet de décrire la langue de manière plus fine par exemple `fr-CH` désigne le français tel qu'il est parlé en Suisse.

Il y a d'autres caractéristiques que la langue ou le pays. Ainsi, `sr-Latn-CS` représente le serbe (`sr`) écrit dans l'alphabet latin (`Latn`) tel qu'il s'utilise en Serbie (`CS`).

La question étant sensible (le croate est-il une langue différente du serbe, par exemple ?) l'IETF a évité les problèmes en s'appuyant sur des normes existantes (ISO 639 pour les langues comme le RFC 1591 s'appuie sur ISO 3166 pour éviter l'épineuse question de « qu'est-ce qui est un pays »). Néanmoins, le RFC confie un rôle de registre à l'IANA pour garantir une stabilité des noms (l'ISO ne la garantit pas, ne s'intéressant qu'au présent, alors que l'Internet a besoin de stabilité sur le long terme, cf. la section 3.4 pour la politique de stabilité du registre des langues).

ISO 639-1 et 2 reflétaient plutôt les besoins des bibliothécaires, soucieux de simplifier la classification en évitant la multiplication des langues. ISO 639-3 adopte plutôt le point de vue des linguistes, qui tendent à voir bien plus de langues. Ce débat entre "*mergers*" (les bibliothécaires) et "*splitters*" (les linguistes) ne cessera pas de sitôt. L'intégration de ISO 639-3 <`https://www.bortzmeyer.org/iso-639-3.html`> fait que le registre des étiquettes de langues passe plutôt du côté des "*splitters*".

Les étiquettes de langue sont donc composées de **sous-étiquettes**, et les diverses restrictions qui existent sur leur longueur et leur place font qu'il est possible d'analyser une étiquette, de déterminer où est la langue, où est le pays, etc, sans avoir d'accès au registre (section 2.2). Une étiquette peut être **bien formée** (obéir à la syntaxe) même si ses sous-étiquettes ne sont pas enregistrées (section 2.2.9 pour la définition de la conformance à cette norme, et la notion de « bien formée » et de « valide »). Le registre stocke des sous-étiquettes, pas des étiquettes entières, et ces sous-étiquettes peuvent être combinées librement, même si le résultat n'est pas forcément sensé. Par exemple, `ar-Cyrl-AQ` est à la fois bien formée - elle obéit à la grammaire, valide (toutes ses sous-étiquettes sont enregistrées) et néanmoins inutile car désignant l'arabe en écriture cyrillique tel que pratiqué en Antarctique.

Les différentes parties d'une étiquette sont décrites en section 2.2. Par exemple, l'écriture est complètement traitée en section 2.2.3 est dérive d'ISO 15924.

La section 2.2.6, sur les extensions (qu'il ne faut pas confondre avec les "*extlangs*") n'est pas encore utilisée ; pour l'instant, aucune extension n'a été enregistrée.

Le format et la maintenance du registre sont décrits en section 3. Le registre est au format "*record-jar*" (section 3.1.1, ce format est décrit dans le livre "*The Art of Unix programming*" <`https://www.bortzmeyer.org/art-unix-programming.html`>) et, par exemple, l'enregistrement du français est :

```
Type: language
Subtag: fr
Description: French
Added: 2005-10-16
Suppress-Script: Latn
```

et celui du cantonais est :

```
Type: language
Subtag: yue
Description: Yue Chinese
Added: 2009-07-29
Macrolanguage: zh
```

et cet enregistrement comporte la mention de la macrolangue (le chinois).

Désormais, le registre est en UTF-8 ce qui permet des choses comme :

```
Type: language
Subtag: pro
Description: Old Provençal (to 1500)
Description: Old Occitan (to 1500)
Added: 2005-10-16
```

L'ajout d'éléments au registre est possible, via une procédure décrite en sections 3.5 et 3.6. Si vous voulez enregistrer une sous-étiquette, la lecture recommandée est « "*How to register a new subtag*" <http://www.langtag.net/register-new-subtag.html> ». La procédure inclus un examen par un "*Language Subtag Reviewer*" (section 3.2), actuellement Michael Everson. Je préviens tout de suite : comme indiqué en section 3.6, la probabilité qu'une demande refusée par l'ISO soit acceptée dans le registre IANA est très faible (malgré cela, des demandes sont régulièrement reçues pour des « langues » que seuls certains radicaux considèrent comme telles). Une liste complète des demandes d'enregistrement effectuées est archivée en <https://www.iana.org/assignments/lang-subtags-templates>. Un exemple réel est décrit dans « Enregistrement de l'alsacien dans le registre IETF/IANA <https://www.bortzmeyer.org/enregistrement-alsacien.html> ».

Lorsqu'on est occupé à étiqueter des textes, par exemple à mettre des attributs `xml:lang` dans un document XML, on se pose parfois des questions comme « Dois-je étiqueter ce texte comme `fr` ou bien `fr-FR`, lorsque les deux sont légaux ? » La réponse figure dans la section 4.1, qui est résumée dans « "*Tag wisely*" <http://www.langtag.net/tag-wisely.html> » (la réponse, ici, est « En général, `fr` tout court, sauf s'il est important de le distinguer des dialectes étrangers à la France »).

Les changements les plus importants par rapport au RFC 4646 (section 8) sont :

— Intégration des normes ISO 639-3 <https://www.bortzmeyer.org/iso-639-3.html> et 5 <https://www.bortzmeyer.org/iso-639-5.html> (section 2.2.1) ce qui fait passer le nombre de langues de 500 à 7000.
— L'arrivée de ISO 639-3 a apporté les extlangs <https://www.bortzmeyer.org/extlang-or-not-extlang.html> ("*Extended Language Subtag*", section 2.2.2) et 639-5 les collections, groupes de langues reliées entre elles.
— Meilleure description du rôle de l'IANA, surtout pour l'archivage des requêtes.
— Quelques changements pour les implémenteurs : grammaire différente mais meilleure (section 2.1) (notamment par une meilleure séparation des aspects syntaxiques et sémantiques) et registre désormais en UTF-8 et plus en ASCII.
— Et plein de petits détails.

Le changement de grammaire ne change pas grand'chose aux étiquettes bien formées ou pas mais il facilitera la tâche des implémenteurs. Les étiquettes patrimoniales ("*grandfathered*") sont ainsi séparées en deux catégories, selon qu'elles obéissent quand même à la syntaxe standard ou pas.

L'augmentation importante de la taille du registre fait que, encore plus qu'avant, la récupération automatique du registre nécessite donc de faire attention pour ne pas charger les serveurs de l'IANA (la

méthode recommandée est le `GET` conditionnel de HTTP RFC 7232 (`If-Modified-Since`, section 3.3), par exemple, avec curl, `curl --time-cond "20090702" https://www.iana.org/assignments/language`)

Les extlangs ont été une question particulièrement délicate. Parmi les macrolangues (une collection de langues qui, dans certaines cicronstances, peuvent être vues comme une langue unique), le piège spécifique à l'arabe et au chinois est que le mot « chinois » désigne à la fois une macrolangue **et** une langue particulière couverte par cette macrolangue (le mandarin). Pire, dans les deux cas (chinois et arabe), les communautés locales ont du mal à admettre le verdict des linguistes. Ceux-ci disent que l'« arabe » n'est pas une langue, et que le marocain et le syrien sont des langues distinctes. Les arabophones ne sont pas tous d'accord avec l'analyse scientifique.

Les autres macrolangues n'ont pas de « langue de référence » et elles sont typiquement pratiquées par des communautés ayant moins de poids que le gouvernement de Pékin.

L'IETF a tranché que, pour certaines langues, l'"*extlang*" serait PERMIS, la forme sans "*extlang*" RECOMMANDÉE (voir notamment la section 4.1.2 sur l'usage des "*extlangs*"). Les vainqueurs sont donc : 'ar' (Arabe), 'kok' (Konkanî), 'ms' (Malais), 'sw' (Swahili), 'uz' (Ouzbèke), et 'zh' (Chinois).

Donc :
— pour le cantonais, `zh-cmn` permis, `cmn` recommandé,
— idem pour l'arabe, `ar-arz` est permis pour l'égyptien, `arz` est recommandé,
— pour le Cri, il n'a pas bénéficié de l'exception donc le Cri des Plaines est forcément `crk`, **pas** `cr-crk`.

Les transparents d'un exposé sur ces "*language tags*" sont disponibles (en ligne sur `https://www.bortzmeyer.org/files/langtags-PRINT.pdf`). Ils ne parlent que de la version précédente, celle du RFC 4646. Le site Web officiel sur les étiquettes de langues est <http://www.langtag.net/>.

Un analyseur de "*language tags*" en logiciel libre existe, GaBuZoMeu <https://www.bortzmeyer.org/gabuzomeu-parsing-language-tags.html> et gère cette nouvelle norme. On peut trouver d'autres mises en œuvre en <http://www.langtag.net/>.