RFC 4646 : Tags for Identifying Languages

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 11 septembre 2006. Dernière mise à jour le 10 octobre 2006

Date de publication du RFC : Septembre 2006

http://www.bortzmeyer.org/4646.html

Successeur du très utilisé RFC 3066[1], puis lui-même remplacé par le RFC 5646, 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 2616), 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 en général deux 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/rfc3066.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).

Les changement par rapport au précédent, le RFC 3066 sont détaillés à la fin du RFC. Le changement plus visible est le choix d'avoir désormais un registre complet à l'IANA. L'ancien registre IANA était très incomplet (beaucoup de langues n'étaient pas enregistrées) et ne séparait pas clairement les étiquettes ("*tag*", la langue) des sous-étiquettes ("*subtag*", le pays et/ou l'écriture), le "*tag*" complet étant considéré comme opaque aux applications. Le nouveau registre, initialisé par le RFC 4645 est bien plus complet. Il s'appuie toujours sur les normes existantes (ISO 639 pour les langues, ISO 15924 pour les écritures et ISO 3166 pour les pays). La syntaxe, si elle ne change pas en apparence (`fr-BE` sera toujours une étiquette valable pour le français pratiqué en Belgique), est plus rigoureuse. À son tour, notre RFC 4646, a été remplacé par un RFC plus récent, le RFC 5646.

Les auteurs du RFC ont expliqué leurs choix dans les excellents articles "*Reasons for Enhancing RFC 3066*" <http://www.inter-locale.com/ID/why-rfc3066bis.html> et "*Understanding the New Language Tags*" <http://www.w3.org/International/articles/bcp47/>.

Les transparents d'un exposé sur ces "*language tags*" sont disponibles (en ligne sur `http://www.bortzmeyer.org/files/langtags-PRINT.pdf`).

Un analyseur de "*language tags*" en logiciel libre existe désormais, GaBuZoMeu <http://www.bortzmeyer.org/gabuzomeu-parsing-language-tags.html>.