Les *language tags*

Stéphane Bortzmeyer
AFNIC
bortzmeyer@nic.fr

10 octobre 2006

Exposé libre

Le problème

Sur Internet, on a souvent besoin d'un moyen de désigner des langues.

Deux cas typiques :

1. Méta-donnée : indiquer la langue utilisée dans un document (XML, par exemple) pour la recherche, le formatage selon des règles spécifiques à une langue, etc.
2. Indiquer à un serveur la langue préférée de l'utilisateur (ce que font les navigateurs Web).

L'approche

Il faut donc un mini-langage normalisé pour décrire les langues.

Le problème est complexe car il y a aussi :

- les variantes nationales,
- les différentes écritures, etc.

Ne pas manquer `http://www.ethnologue.com/`.

Quelques cas rigolos

1. L'azéri, langue turque, s'est d'abord écrit en alphabet arabe, puis cyrillique, puis latin.
2. Le français n'est pas tout à fait le même à Abidjan, Tunis ou Bruxelles.
3. Le "chinois" n'existe pas vraiment : ce mot recouvre une famille de langues.
4. La réforme orthographique de l'allemand en 1996 n'a pas été adoptée par la Suisse. Même en Allemagne, tout le monde ne la suit pas.
5. Outre les différences selon le pays, il y a des dialectes différents au sein d'un même pays.
6. Le serbe et le croate sont-ils deux langues différentes ? **Attention** : question brûlante.

Le contexte

Il existe plusieurs normes sur ce sujet :

- ISO 639 : une famille de normes ISO qui donne des codes aux **langues**. `fr` est le français, par exemple. Dans ISO 639-1, les codes ont deux caractères et dans ISO 639-2, ils en ont trois (`fra` pour le français, `apa` pour l'apache, `jbo` pour le lojban).
- ISO 3166-1 : une norme ISO qui donne des codes de deux (et de trois) caractères aux **pays**. **Ne pas confondre** le `fr` de 3166-1 (la France) avec le `fr` de 639-1 (le français).
- ISO 15924 : une norme ISO qui donne des codes de quatre caractères aux **écritures** comme `Latn` pour le latin ou `Cyrl` pour le cyrillique.
- UN M.49 : une norme de l'ONU (reprise, sauf erreur, dans ISO 3166-1) pour les régions multi-nationales. Les codes M.49 ont trois chiffres (151 : "Europe de l'Est", 061 : "Polynésie").
- Unicode n'est pas concerné ici.

Stabilité des normes ISO

Attention, l'ISO ne garantit aucune stabilité. Ces normes ne concernent que le présent.

`CS` a disparu de ISO 3166 avec la Tchécoslovaquie, avant d'être réattribué à la Serbie-et-Monténégro.

L'ancienne solution

Le RFC 3066 (qui remplaçait le 1766) a normalisé la syntaxe dite *language tag*. Un *tag* comprenait :

1. Un code de langue, tiré de ISO 639-1,
2. Des *subtags* optionnels, notamment les codes de pays tirés de ISO 3166,
3. Quelques cas particuliers comme les codes IANA, commençant par "i-" (`i-klingon`).

Exemples : `fr` (français) ou `fr-CA` (français tel que parlé au Canada).

S'appuyer sur ISO permettait d'éviter les problèmes politiques (serbe et croate, moldave et roumain, l'andalou est-il une langue ou un dialecte du castillan, etc).

RFC 3066

Ce RFC a été un grand succès. Le protocole HTTP a été son principal utilisateur (champ `Accept-Language` du protocole).

Plusieurs normes non-IETF s'appuient sur ce RFC, la plus connue étant XML.

In document processing, it is often useful to identify the natural or formal language in which the content is written. A special attribute named `xml:lang` *may be inserted in documents to specify the language used in the contents and attribute values of any element in an XML document. The values of the attribute are language identifiers as defined by [IETF RFC 3066], Tags for the Identification of Languages, or its successor ;*

Les limites du RFC 3066

- Manque de stabilité des normes ISO sous-jacentes. L'Internet nécessite que `cs-CS`, une fois valide, reste valide même si la Tchécoslovaquie disparait.
- Fermeture de l'ISO, dont les normes ne sont pas disponibles, au sens RFC du terme.
- Impossibilité d'identifier l'écriture (le pays était parfois utilisé à cette fin, par exemple l'écriture chinoise simplifiée est utilisée surtout en Chine et l'écriture chinoise traditionnelle surtout à Taïwan).
- Et le langage d'expression des *tags* manquait de souplesse.

Le cahier des charges

- Tout *tag* valide avec le RFC 3066 devait rester valide (même ceux qu'on regrettait d'avoir enregistrés).
- Le registre devait être stable et ouvert.

La nouvelle solution

Le groupe de travail LTRU (*Language Tag Registry Update*) a travaillé sur ce cahier des charges.

Le résultat est le RFC 4646, paru en septembre 2006.

Il reprend cette syntaxe mais coupe partiellement les ponts avec l'ISO en créant un registre IANA des langues. Cette scission a fait couler beaucoup d'encre.

Il permet en outre de spécifier des cas bien plus complexes.

Les nouveaux *tags*

- L'écriture s'insère entre la langue et le pays : `az-Arab-IR` (azéri en caractères arabes, tel que parlé en Iran).
- Des nouveautés dans le langage (comme les extensions).

La nouvelle grammaire est **générative** : tout ensemble de *subtags* valide est valide. Il n'y a pas besoin d'enregistrer le *tag*.

Quelques exemples

- `mn` : le mongol (sans autres précisions)
- `en-Arab` : l'anglais écrit en caractères arabes (inhabituel mais valide)
- `en-FR` : l'anglais tel que parlé en France
- `fr-029` : le français, tel que parlé dans les Caraïbes
- `uk-Latn-MD` : l'ukrainien parlé en Moldavie, écrit en caractères latins.

Chercher un *tag* qui convient

Le RFC 4647, sur le *matching* des *tags* spécifie deux cas :

1. Celui où on cherche tous les *tags* correspondant à un certain critère. Exemple : tous les documents en finnois (`fi`) dans une base.
2. Celui où on cherche **un** document correspondant le mieux (ou le moins mal) à l'utilisateur. Exemple : un serveur Web qui doit choisir la page à renvoyer.

L'algorithme est esquissé dans les deux cas, mais pas spécifié complètement (car cela dépend de l'application).

Correspondance, suite

Dans les deux cas, le client peut spécifier un **intervalle** (*language range*) de langues. Exemple en HTTP :

```
Accept-Language : da, en-GB;q=0.8, en;q=0.7
```

(Danois, sinon anglais britannique, sinon anglais quelconque.)

L'intervalle peut comporter des jokers (*).

Technique : la grammaire

Pour être compatible avec le RFC 3066 et pour pouvoir être analysée relativement facilement, la grammaire limite le nombre de caractères de chaque `subtag`.

Exemple (simplifié) : entre `man-Nkoo` (mandingue écrit en N'ko) et `man-GN` (mandingue parlé en Guinée), c'est la longueur des deux *subtags* qui permet de savoir ce qui est un pays et ce qui est une écriture.

L'accès au registre IANA n'est donc pas indispensable.

Il ne sert que si on veut tester la **validité** d'un *tag*. Comme en XML, un *tag* doit être **bien formé** mais n'est pas forcément **valide**.

Technique : une mise en œuvre

GaBuZoMeu est une mise en œuvre du RFC en logiciel libre. Il permet de :

- Vérifier qu'un *tag* est bien formé,
- Vérifier qu'un *tag* est valide,
- Afficher les *subtags*,
- Traduire le registre en XML ou en CSV,
- ...

```
% check-wf en-Arab-SA
en-Arab-SA is well-formed
% check-wf en-Ar-SA
en-Ar-SA is NOT well-formed: ...
% check-valid en-Arab-SA
en-Arab-SA is valid
% check-valid en-Lati-SA
en-Lati-SA is NOT valid: Unknown script Lati
% display-tag en-Arab-SA
en-Arab-SA: (en: language "English", Arab: script  "Arabic", SA: region  "Sau
```

Futur

Le travail sur RFC 4646 bis a déjà commencé. Gros chantier : intégration de la future ISO 639-3, qui va apporter :

- Dix fois plus de langues,
- Les concepts de collections et de **macrolangues**.

Macrolangues

Une macrolangue est un groupe de langues distinctes, mais qui ont parfois besoin d'être manipulées ensemble. Exemple traditionnel : le chinois, qui regroupe, le cantonais, le mandarin, etc.

Autre exemple, `oci`, l'occitan, qui regroupe `auv` (auvergnat), `gsc` (gascon), `lms` (limousin), `lnc` (languedocien), `prv` (provençal).

Grandes questions : gouvernance

Un problème de gouvernance : qui décide des langues humaines ?

- L'ISO, lente, inefficace, chère et fermée ?
- L'IANA, le délégué du gouvernement états-unien ?
- Le consortium Unicode, organisme privé ?
- L'IETF, où personne n'y connait grand'chose ?
- Les Nations Unies ? :-)

Opinion personnelle

Je considère que l'ISO est le principal responsable de la scission de l'IANA.

- extrême fermeture au public (normes non disponibles gratuitement et facilement),
- grande lenteur de publication (même en comparaison de l'IANA),
- absence de sens pratique (pas de distribution du registre sous une forme analysable par un programme)
- refus de considérer l'importance de la stabilité (les normes ISO comme 3166, les codes de pays, sont une photographie de la situation actuelle du monde, alors que les noms de domaines ont besoin de stabilité : heureusement que l'ICANN n'a pas supprimé "su" de la racine).

Performances de l'ISO

Le Zazaki (zza) a été ajouté à l'ISO 639-2 début 2006. Malgré la revue effectuée par l'IETF et malgré les performances de l'IANA, il a été ajouté au registre le 24 août.

Le 18 septembre, il n'est toujours pas sur le site Web de la *Maintenance Agency* de ISO 639. . . (Il a été ajouté en octobre.)

Grandes questions : Internet et multilinguisme

Pouvoir *taguer* les textes et les requêtes ne va pas créer du contenu multilingue !

L'IETF fournit l'infrastructure logicielle, encore faut-il s'en servir.