Indiquer la langue, l'écriture, le pays dans des documents informatiques

Patrick Andries

Conseils Hapax, Québec, Canada
Membre du consortium Unicode
patrick@hapax.qc.ca

Résumé. Dans cette communication nous verrons comment préciser la langue, le dialecte ainsi que d'autres informations culturelles comme le système d'écriture utilisé ou le pays d'où provient le scripteur, et les appliquer à des données Unicode. Ces informations peuvent se révéler précieuses pour une kyrielle de processus de traitement des documents.
In this paper, we will review how to specify the language, the script and the country of an electronic document, or parts thereof. This information, as we will see, is valuable for a series of automated text processes.
Mots-clés. Unicode, langue, écriture, pays, dialecte, locale, amazighe, informatique, français, chinois, tifinaghe, Unicode, arabe, ISO 639, ISO 3166, RFC 5646, RFC 4646, ISO 15924, BCP 47, IETF.

1 Introduction

Préciser la langue, ainsi que d'autres informations culturelles comme le pays d'où provient le scripteur, et les appliquer à des données Unicode permet :

- de résoudre les ambiguïtés d'affichage (exemple : « désunifier » les caractères CJC),
- de rendre la synthèse vocale possible,
- d'obtenir les bonnes ressources linguistiques d'un programme internationalisé
- d'afficher la bonne version linguistique d'une page internet,
- de mieux effectuer la coupure de lignes et de mots,
- d'obtenir de meilleurs résultats quand on classe, cherche ou trie ces données,
- et de permettre la correction orthographique.

En règle générale, on peut dire que la langue est orthogonale au codage de caractères : connaître le codage d'un texte ne permet pas d'en deviner la langue. C'est ainsi qu'un texte codé en Latin-1 peut très bien être en anglais, en espagnol ou en français, voire les trois à la fois.

Nous allons passer en revue ci-dessous les normes qui établissent la manière de préciser la langue et d'autres métadonnées culturelles d'importance.

2 ISO 639 – indicatifs de langue

L'ISO 639 est une norme internationale qui définit des indicatifs pour la représentation des noms de langues. Elle est actuellement composée de 3 parties. Sa première partie, l'ISO 639-1[1] (alpha-2), utilise des codets ou indicatifs sur 2 caractères, et les associe à des noms de langue en français et en anglais. L'ISO 639-2[2] (alpha-3) utilise des codets sur 3 caractères et connaît deux formes de codages possibles : ISO 639-2/B (bibliographique) et ISO 639-2/T (terminologique). En règle générale, les indicatifs bibliographiques ressemblent à ceux définis par la norme américaine Z39.53 et s'inspirent des noms qui désignent ces langues en anglais, alors que les indicatifs terminologiques ressemblent aux noms que ces langues se donnent. Enfin, l'ISO 639-3[3] complète l'ISO 639-2 ; ces indicatifs à trois lettres sont tirés de la base de données de l'Ethnologue[4].

Tableau 1. Les différentes parties de l'ISO 639

Partie	Type d'indicatif	Nombre d'indicatifs	Exemples
ISO 639-1	à deux lettres	136	« fr » pour le français, « wa » pour le wallon, « ar » pour l'arabe.
ISO 639-2	à trois lettres	484	« fre » et « fra » pour le français, « ber » pour le berbère
ISO 639-3	à trois lettres	7581	« fra » pour le français, « rif » pour le rifain.

[1] Liste non officielle sur <http://fr.wikipedia.org/wiki/Liste_des_codes_ISO_639-1>.
[2] Liste sur <http://www.loc.gov/standards/iso639-2/php/French_list.php>.
[3] Liste sur <http://www.sil.org/ISO639-3/codes.asp>.
[4] Voir <http://www.ethnologue.com/>.

L'ISO 639 n'est pas un registre stable. On y ajoute, de temps à autre, de nouveaux indicatifs et plusieurs langues en ont changé. C'est le cas de l'hébreu, de l'indonésien et du yidiche qui, pour deux d'entre eux, sont passés à des indicatifs qui s'inspirent de leur nom ou graphie en anglais plutôt que dans la langue en question (de « iw » à « he » et de « ji » à « yi »). Les logiciels doivent prendre en charge les deux versions de ces indicatifs et les considérer comme synonymes.

Tableau 2. *Quelques indicatifs tirés de l'ISO 639-1*

Indicatif ISO 639-1	Langue
ar	arabe
de	allemand
el	grec
en	anglais
es	espagnol
fr	français
it	italien
nl	néerlandais
pt	portugais
ru	russe

Vingt-deux langues ont dans l'ISO 639-2 deux codets (indicatifs) différents : l'un bibliographique (ISO 639-2/B) et l'autre terminologique (ISO 639-2/T). C'est le cas du français qui s'écrit « fre » (bibliographique) ou « fra » (terminologique). Il faut considérer, à nouveau, ces indicatifs comme synonymes. En pratique, ces doublons sont rarement utilisés puisque ces langues ont également un indicatif de deux lettres (ISO 639-1) et que les normes et standards préconisent alors l'utilisation de celui-ci.

L'ISO 639-2 prévoit une zone à usage privé qui s'étend de « qaa » à « qtz », l'ISO n'affectera ces codets à aucune langue normalisée. Ces indicatifs peuvent être utilisés par accord commun entre des tiers.

Dans les applications qui utilisent les codets de langue ISO 639, il est préférable d'utiliser en premier lieu un indicatif alpha-2 de l'ISO 639-1, s'il existe. Si ce n'est le cas, on choisira le codet alpha-3 de l'ISO 639-2/T. Enfin, en dernier ressort, on pourra utiliser les indicatifs alpha-3 de la norme ISO 639-3.

2.1 Macrolangue

Certains des indicatifs de l'ISO 639-3 correspondent à des macrolangues. Une macrolangue est un ensemble de langues étroitement apparentées ou de dialectes fortement divergents. On compte 56 langues dans ISO 639-2 qui sont considérées comme des macrolangues dans ISO 639-3. L'arabe (« ar » dans ISO 639-1 et « ara » dans ISO 639-2) est une de ces macrolangues dans l'ISO 639-3 (« ara »). Le chinois est également considéré comme une macrolangue (« zh ») et une des langues de cette macrolangue est « cmn », appelée le mandarin. Notons que, dans le cas de l'arabe et du chinois, les communautés locales ont du mal à admettre le verdict des linguistes. Ces érudits considèrent habituellement que l'arabe n'est pas une langue, mais que le marocain et le syrien sont des langues distinctes. Les arabophones, en revanche, ne partagent pas cette analyse.

Tableau 3. *Les « langues berbères » dans l'ISO 639-3*

Indicatif ISO 639-3	Langue
shi	tachelhit ou chleuh
tzm	tamazight (centre du Maroc)
rif	tarifit ou rifain
cnu	chénoui
shy	tachaouit, chaoui
kab	kabyle
thv	tamachek (Sud algérien)
thz	tamachek (Agadez)

L'ISO 639-1 et 2 reflétaient en gros les besoins des bibliothécaires, soucieux de simplifier la classification en évitant de multiplier les langues. L'ISO 639-3, pour sa part, adopte plutôt le point de vue des

linguistes qui tendent à voir nettement plus de langues. Ce débat entre les « fusionneurs » (les bibliothécaires) et les « diviseurs » (les linguistes) se poursuit. L'intégration de l'ISO 639-3 illustre l'importance des linguistes (et du SIL d'où proviennent ces indicatifs) dans la conception du registre des langues.

Quant à l'ISO 639-5, il définit des codes alpha-3 (à trois lettres) pour les familles ou groupes de langues. Certains de ces codes font également partie de l'ISO 639-2, où se mêlaient aussi des codes pour des macro-langues et langues individuelles, mais pas de façon assez précise ni complète.

Tableau 4. *Exemples de famille et groupes de langues dans l'ISO 639-5*[5]

Indicatif ISO 639-5	Groupe	Hiérarchie
afa	Langues afro-asiatiques	afa
ber	Langues berbères	afa : ber
cdc	Langues tchadiques	afa : cdc
cus	Langues couchitiques	afa : cus
egx	Langues égyptiennes	afa : egx
ine	Langues indo-européennes	ine
itc	Langues italiques	ine : itc
omv	Langues omotiques	afa : omv
roa	Langues romanes	ine : itc : roa
sem	Langues sémitiques	afa : sem

3 ISO 3166 – indicatifs de pays

Les indicatifs de pays normalisés par l'ISO 3166-1[6] forment un deuxième type de renseignement culturel. L'ISO 3166-1 comprend trois répertoires différents : alpha-2, alpha-3 et numérique-3. L'ISO 3166-1

[5] Pour une liste complète : < http://www.loc.gov:8081/standards/iso639-5/fr.php> .
[6] On peut consulter gratuitement la liste officielle complète à l'adresse suivante : <http://www.iso.org/iso/fr/country_codes/iso_3166_code_lists/french_country_names_and_code_elements.htm>

alpha-2 définit une série d'indicatifs de pays sur deux lettres. Alpha-3 précise des codets de pays sur 3 lettres. Il existe également une norme ISO 3166-2, codée à l'aide de quatre lettres, qui permet de désigner des subdivisions de pays.

Tableau 5. *Quelques indicatifs de pays ISO 3166-1 à deux lettres*

Pays	Indicatif
Algérie	DZ
Belgique	BE
Canada	CA
Égypte	EG
Espagne	ES
France	FR
Libye	LY
Mali	ML
Malte	MT
Maroc	MA
Mauritanie	MR
Niger	NE
Tchad	TD
Tunisie	TN

Les codets ISO 3166-1 (sur deux lettres) correspondent habituellement aux noms de domaines de tête par pays[7] (en anglais, « Country Coded Top Level Domains », abrégé en *ccTLD*) utilisés dans l'attribution des adresses internet (voir le « ma » dans « ircam.ma »). Il existe quelques exceptions dont la plus notable est sans doute celle du Royaume-Uni qui a un indicatif 3166 égal à « GB » (Grande-Bretagne), mais dont le domaine internet est « .uk » (« United Kingdom », Royaume-Uni).

On remarque que, à l'instar de l'ISO 639, l'ISO 3166 n'est pas un ensemble figé d'indicatifs, certains pays naissent (par exemple à la suite de l'éclatement de l'Union soviétique), d'autres changent de nom et

[7] Pour plus de détails sur ces termes techniques, voir l'autre communication de ces mêmes actes : *Demain, encore plus de tifinaghes sur Internet.*

voient leur indicatif tomber en désuétude ou être repris par un autre pays. C'est le cas des Territoires des Afars et Issas, aujourd'hui Djibouti, dont le codet AI a été repris par l'île d'Anguilla. Il en va de même du codet GE, anciennement attribué aux îles Gilbert et Ellice devenues Tuvalu, et maintenant utilisé par la Géorgie. On imagine facilement les problèmes de données historiques qui utiliseraient les codets 3166 pour indiquer un pays. Ce genre de complication est d'ailleurs à la source d'une récente norme internet sur laquelle nous reviendrons bientôt : le RFC 5646.

La norme **ISO 3166-2** (seconde partie de la norme ISO 3166), édictée par l'Organisation internationale de normalisation, permet de désigner les principales subdivisions administratives d'un pays par un codet en quelques chiffres ou lettres complétant le code ISO 3166-1 du pays.

Tableau 6. *Quelques codets de l'ISO 3166-1 et de l'ISO 3166-2*

Indicatif	Signification
FR-01	Département de l'Ain en France
MA-CAS	Province de Casablanca
MA-OUA	Province d'Ouarzazate
MA-TET	Province de Tétouan
MA-01	Région Tanger-Tétouan
MA-06	Région de Meknès-Tafilalet

4 M.49 – Indicatifs de pays et de régions

Les codets ONU M. 49 sont des indicatifs à trois chiffres attribués par la division de la statistique de l'ONU. Ils ne correspondent pas toujours à des pays ; le codet 001, par exemple, représente le monde entier, 002 l'Afrique et 015 l'Afrique septentrionale. Le codet à trois chiffres d'un pays défini dans l'ISO 3166-1 numérique-3 est identique au codet M.49 défini pour le même pays. Toutefois, certains indicatifs M.49 n'ont pas de correspondance ISO 3166-1 quand ils représentent une région supranationale ou infranationale[8].

[8] Liste complète officielle ici : <http://unstats.un.org/unsd/methods/m49/m49alphaf.htm>.

Tableau 7. *Quelques indicatifs M.49*

Indicatif	Signification
012	Algérie
434	Libye
466	Mali
478	Mauritanie
504	Maroc
562	Niger
788	Tunisie

5 ISO 15924 – indicatifs d'écriture

L'ISO 15924 définit un indicatif pour près de 150 écritures différentes. Nous en reproduisons quelques-uns dans le tableau 4[9]. Rappelons que le français, par exemple, relève de l'écriture latine (Latn) et l'amazighe normalisé au Maroc du tifinaghe (Tfng).

Tableau 8. *Quelques indicatifs d'écriture*

Indicatif	Signification
Arab	Arabe
Copt	Copte
Cyrl	Cyrillique
Grek	Grec
Hani	Idéogrammes japonais
Hans	Idéogrammes chinois simplifiés
Hant	Idéogrammes chinois traditionnels
Latf	Latin brisé (« fraktur » ou gothique)
Latn	Latin

[9] La liste complète peut être consultée à l'adresse suivante : <http://www.unicode.org/iso15924/iso15924-fr.html>

Indicatif	Signification
Tfng	Tifinagh (tifinar).

6 RFC 5646 – étiquettes de langue

L'Internet Engineering Task Force, en abrégé IETF, littéralement le « Détachement d'ingénierie d'Internet », est un regroupement informel d'experts qui élabore les standards de l'Internet. Les *Request for comment* (RFC, les *demandes de commentaires*) sont une série de documents portant sur l'Internet émis par l'IETF. Peu de ces RFC sont des standards, mais tous les standards de l'Internet sont enregistrés en tant que RFC. Chaque RFC porte un numéro séquentiel unique. Une fois adopté, un RFC n'est jamais modifié ou retiré. Si un RFC doit être modifié, on publie un autre RFC (avec un autre numéro) qui le remplace.

C'est ce qui s'est produit en 2009 quand le RFC 5646 a complété et remplacé le RFC 4646 qui définissait déjà des étiquettes linguistiques. La plus grande innovation du RFC 5646 est l'introduction des milliers de codets de l'ISO 639-3 et ceux de l'ISO 639-5 qui définit les ensembles de langues (par exemple « afa » pour « les langues afro-asiatiques »). Son prédécesseur, le RFC 4646, remplaçait déjà un standard précédent sur le même sujet, le RFC 3066 qui lui-même remplaçait le RFC 1766. Toutes les étiquettes de langue définies selon le RFC 1766 restent conformes aux RFC 4646 et 5646. Les étiquettes de langue du RFC 5646 consistent en une série de sous-étiquettes séparées par des traits d'union, sans égard à la casse. Voici un aperçu des règles de construction des étiquettes définies par le RFC 5646 (nous expliquerons cette « grammaire » ci-dessous) :

```
Une étiquette de langue consiste en
                étiqlang            ; étiquette générative
            -ou- usage-privé        ; une étiquette privée
            -ou- patrimonial        ; anciennes valeurs
                                    ; enregistrées

étiqlang     = (langue
                ["-" écriture]
                ["-" région]
                ("-" variante)*
                ("-" extension)*
                ["-" usage privé])
```

```
langue =   [a-wA-W]{2,3}        ; codet ISO 639 le plus court
           ["-" extlang]        ; suivi d'un extlang optionnel ou
         | [a-wA-W]{4}          ; pour normalisation ultérieure[10] ou
         | [a-wA-W]{5,8}        ; une valeur enregistrée auprès de
                                ; l'IANA[11].

extlang = [a-wA-W]{3}         ; jusqu'à 3 codets ISO 639 séparés
                              ; par un "-", en pratique un seul
                              ; codet est utilisé

écriture     = "Latn", "Cyrl" ; codets ISO 15924 (4 lettres)

région       = "US", "CS", "FR"... ; codets ISO 3166 à 2 lettres
               "419", "019"...     ; codets ONU M.49 à 3 chiffres

variante     = "rozaj", "1996"...  ;  plusieurs sous-étiquettes
                                   ; permises de 4 à 8
                                   ; alphanumériques

extension    = une lettre suivie de sous-étiquettes, plusieurs
               extensions sont permises pour une même étiquette de
               langue.

usage-privé  = "x-" suivi de sous-étiquettes, autant que
               nécessaires, peuvent être en début ou en fin
               d'étiquette, mais pas au milieu.

patrimonial  = étiquettes reprises de l'ancien registre (RFC 3066)
               et qui ne sont pas des doublons (une liste fermée).
```

Chaque type de sous-étiquette a une longueur précise et des restrictions quant à son contenu. L'étiquette commence toujours par la sous-étiquette « langue » qui peut être un codet ISO 639 ou une autre valeur enregistrée auprès de l'IANA[12]. Elle peut être suivie de différentes sous-étiquettes. Il existe à l'heure actuelle cinq types de sous-étiquettes qui peuvent suivre l'indicatif de langue : l'écriture, la région, les variantes, les extensions et l'usage privé. L'ordre, la longueur et le contenu de chaque sous-étiquette sont bien établis.

Toutes les sous-étiquettes légitimes sont consignées dans un registre unique tenu à jour par l'IANA, plutôt que les différentes agences de mise à jour des normes ISO comme c'était le cas pour le RFC 3066. Pour la sous-étiquette « langue », l'IANA n'enregistre qu'un seul codet par

[10] Réservé pour faire face à l'épuisement possible des codets ISO 639 à 3 lettres.
[11] Pour les très rares cas où l'IANA accepterait d'enregistrer une langue refusée par l'ISO 639. À l'heure actuelle, aucune sous-étiquette de ce type n'a été attribuée.
[12] L'IANA est un organisme américain responsable de la gestion de l'espace d'adressage IP d'Internet et d'autres ressources requises par les protocoles de communication sur Internet.

langue, alors que l'ISO pouvait en avoir normalisé plusieurs (trois pour le français par exemple : « fr », « fra » et « fre »). Si un codet ISO à deux lettres est disponible, celui-ci apparaîtra dans le registre plutôt que le codet à 3 lettres. Voici l'entrée pour la sous-étiquette de langue « es » dans le registre de l'IANA[13] :

```
%%
Type: language
Subtag: es
Description: Spanish
Description: Castilian
Added: 2005-10-16
Suppress-Script: Latn
%%
```

La ligne « Suppress-Script » indique qu'il ne faut pas mentionner l'écriture de cette langue quand elle est égale à « Latn ». Ceci afin de décourager la prolifération d'étiquettes pléonastiques comme « es-Latn ». Le « Suppress-Script » est également utile pour éviter que de nombreux anciens analyseurs d'étiquettes de langue conçus pour le RFC 3066 dont la syntaxe était, *grosso modo*, « langue » ou « langue-région » ne se plantent quand ils sont confrontés à des étiquettes de langues inutiles. C'est le cas, par exemple, pour les documents à l'étiquette « fr-CA » (français du Canada) valable déjà avec le RFC 3066 et qu'il est inutile de changer en « fr-Latn-CA » car cela pourrait causer des problèmes aux anciens analyseurs.

6.1. Extlang

L'indicatif de langue ISO 639 peut être suivi d'un « extlang » après un « - ». Le premier indicatif est la sous-étiquette principale de langue. L'extlang, s'il est présent, doit précéder toute autre sous-étiquette comme l'écriture ou la région.

Exemples d'étiquette linguistique qui utilisent une sous-étiquette extlang :

- `zh-yue` (chinois cantonais)
- `ar-ary` (arabe marocain)

[13] Le registre se retrouve ici : <http://www.iana.org/assignments/language-subtag-registry>.

Le but principal des combinaisons `<langue principale>-<extlang>` est de prendre en charge des formes historiques d'étiquette linguistique avant, notamment, l'inclusion des milliers de codets de l'ISO 639-3 dans le RFC 5646. Pour chaque forme `<langue principale>-<extlang>`, il existe une forme équivalente qui n'utilise qu'un indicatif `<langue principale>` sans extlang. Cette forme simple est recommandée et il faut l'utiliser dans la mesure du possible. C'est ainsi qu'on préfèrera `yue` à `zh-yue` pour le cantonais et `ary` à `ar-ary` pour l'arabe marocain.

Voici l'entrée du registre qui correspond à l'extlang pour l'arabe algérien du Sahara (`aao`) :

```
%%
Type: extlang
Subtag: aao
Description: Algerian Saharan Arabic
Added: 2009-07-29
Preferred-Value: aao
Prefix: ar
Macrolanguage: ar
```

La `<langue principale>` utilisée avec un extlang est une macrolangue[14] qui comprend un certain nombre de langues ou de dialectes fortement divergents. La sous-étiquette de la macrolangue peut s'utiliser seule, toutefois si son sens n'est pas suffisant clair, le lecteur confronté à un tel document pourrait bien ne pas nécessairement pouvoir le lire.

C'est ainsi que `zh` signifie « chinois», un concept qui recouvre de nombreux « dialectes »[15] qui ne sont pas nécessairement compréhensibles entre eux. Quand on utilise « `zh` » sans plus, on fait le plus souvent référence à la variante dominante dans l'ensemble chinois, à savoir le mandarin (`cmn`), bien qu'il s'agit là d'une convention tacite sur laquelle BCP 47 ne dit mot.

[14] Il existe des macrolangues dont les langues constitutives ne peuvent servir d'extlang, c'est le cas du cri au Canada (la macrolangue) et le cri des Plaines ou le cri de Moose (les langues englobées par la macrolangue). On pourra écrire « cr » (le cri macrolangue), « crk » (cri des Plaines), mais pas « cr-crk ».

[15] Pour les Chinois, le cantonais est par exemple un dialecte. Bien qu'en réalité, d'un point de vue linguistique, il y a plus de différences entre le cantonais et le mandarin qu'entre l'italien et le français, même si l'intercompréhension à l'écrit entre ces deux langues chinoises est assez bonne grâce aux idéogrammes.

Pourquoi les formats avec et sans extlang sont-ils permis ? La principale raison, comme nous l'avons évoqué, est que le RFC 4646, le prédécesseur du RFC 5646, contenait déjà des étiquettes du type `zh-yue` où `zh` était la langue principale et `yue` désignait une variante. C'était une des seules manières[16] de désigner le cantonais qui ne bénéficiait d'aucun indicatif ni dans l'ISO 639-2 ni dans l'ISO 639-1. Avec l'adjonction de l'ISO 639-3 lors de l'adoption du RFC5656, il existe désormais un codet précis qui désigne directement le cantonais (`yue`) qui peut désormais servir dans la sous-étiquette de `<langue principale>`.

Bien que les formes sans extlang soient recommandées, les formes utilisant la macrolangue (avec ou sans extlang) sont toutefois permises et il existe des circonstances où ce choix est sans doute le plus opportun.

C'est le cas, notamment, pour des données ou des applications qui utilisent déjà la sous-étiquette « `ar` » (la macrolangue arabe) et qui préfèreront sans doute continuer d'utiliser cet indicatif plutôt que le nouvel indicatif plus précis « `arb` » (arabe standard moderne).

Notons enfin que le modèle extlang est bien adapté à la négociation de langues (HTTP) où la recherche d'un document en « `ar-ary` » (arabe-arabe marocain) fournira un document identifié comme simplement en arabe (`ar`), solution probablement acceptable si aucun document équivalent en arabe marocain (`ary`) n'est disponible. Notons que l'emploi de cette technique de recherche de document par troncatures successives à droite peut fournir un résultat incompréhensible en fonction de la macrolangue utilisée comme langue primaire (l'indicatif le plus à gauche).

6.2. Valeurs à ne pas utiliser

La plupart des valeurs « patrimoniales » incluses pour des raisons historiques et celles dites redondantes comprennent un champ qui indique qu'elles ne doivent plus être utilisées et qu'on leur préfère désormais un autre indicatif. L'entrée du registre ci-dessous est de ce type.

[16] On utilisait aussi zh-HK par exemple pour désigner la macrolangue chinoise de Hong Kong, où le cantonais est dominant. Cette étiquette est toujours valable, même si elle n'est pas conseillée.

```
%%
Type: grandfathered
Tag:i-lux
Description: Luxembourgeois
Preferred-Value: lb
Deprecated: 1998-09-09
Comments: replaced by ISO code lb
```

Elle peut se lire ainsi : le codet « i-lux » pour le luxembourgeois est de type patrimonial (*grandfathered*), il faut éviter de l'utiliser, on lui préfère la valeur « lb » depuis le 9 septembre 1998.

6.3. Règle d'or

La règle d'or quand on crée des étiquettes linguistiques se résume à toujours utiliser la forme la plus courte possible. Il faut éviter de préciser les sous-étiquettes comme la région, l'écriture ou les autres sous-étiquettes facultatives, sauf si elles apportent une information utile. Ainsi faut-il utiliser « ja » pour le japonais et non « ja-JP » à moins que vous teniez absolument à dire qu'il s'agit du japonais parlé au Japon... Le tableau 9.5 énumère quelques étiquettes linguistiques.

Tableau 9. *Quelques étiquettes linguistiques*

Étiquette	Explications
fr	Le français.
ja	Le japonais.
i-enochian	Exemple d'étiquette patrimoniale : l'énochien[17].
zh-Hant	Le chinois en écriture chinoise traditionnelle.
zh-Hans	Le chinois écrit à l'aide de l'écriture chinoise simplifiée.
sr-Cyrl	Le serbe en cyrillique.
sr-Latn	Le serbe en latin.
sl-Latn-IT-nedis	Le dialecte slovène de Nadiza écrit en latin tel qu'utilisé en Italie (étiquette non recommandée puisque le « Latn »

[17] L'énochien ou « langue des anges » est une langue occulte proposée par les alchimistes anglais John Dee et Edward Kelley au XVI^{e} siècle. Il possède son propre alphabet qu'Unicode n'a pas inclus (on le considère comme une simple transposition de l'alphabet anglais). Le codet i-enochian est un indicatif linguistique, pas un indicatif d'écriture.

Étiquette	Explications
	est redondant pour le slovène, en d'autres termes l'entrée pour « sl » dans le registre de l'IANA a un « Suppress-Script » pour « Latn »).
sl-IT-nedis	Comme ci-dessus, mais sans l'écriture redondante et déconseillée. Cette étiquette est donc meilleure que celle ci-dessus.
de-CH-1901	L'allemand en Suisse écrit avec la variante orthographique de 1901.
zh-yue-Hant-HK	Le cantonais en écriture chinoise traditionnelle à Hong Kong (« zh-yue » est ici un code patrimonial).

6.4. Stabilité garantie

Les changements les plus importants par rapport au précurseur des RFC 5646 et 4646, c'est-à-dire le RFC 3066, ont trait au fait que la syntaxe du RFC 5646 est désormais plus rigoureuse que celle du RFC 3066, que l'IANA tient à jour de manière permanente, stable et gratuite un seul registre unifié et qu'aucun codet ne disparaîtra ou ne sera affecté à une autre entité. L'Internet impose que « cs-CS », une fois valide, reste valide bien que la Tchécoslovaquie ait disparu, en tant qu'entité politique distincte, après l'enregistrement des codets en question. Le registre de l'IANA continue de suivre les normes de l'ISO mentionnées ci-dessus, mais il ne supprime jamais une sous-entrée et des règles claires ont été établies si des conflits entre ce registre et les listes de l'ISO venaient à surgir.

7 RFC 4647, trouver un document dans une langue

Dernier élément de cette série de normes et standards, le RFC 4647[18]. Ce document décrit une syntaxe appelée un *choix de langues* pour construire la liste des préférences linguistiques d'un utilisateur. En d'autres termes, comment préciser que l'on préfère avoir d'abord des

[18] Cf. <http://abcdrfc.free.fr/rfc-vf/rfc4647.htm>

documents en français de France, puis n'importe quelle variante française d'un document équivalent, puis les documents correspondants en arabe.

Ce RFC décrit également plusieurs mécanismes pour comparer ces listes de choix de langue et les associer à des étiquettes de langues. Deux types de mécanismes de correspondance sont définis : le filtrage et la consultation. Le filtrage produit un ensemble (éventuellement vide) d'étiquettes de langues tandis que la consultation produit une seule étiquette de langue. Les applications possibles comprennent la négociation de langue ou la sélection de contenu.

7.1. Filtrage

Dans le cas du filtrage, on cherche à trouver toutes les étiquettes linguistiques qui correspondent à un critère : trouver tous les documents en finnois (`fi`) par exemple.

L'utilisateur précise la valeur la plus générale qui constitue une réponse correcte, c'est ainsi que « `de` » (allemand) correspond à :

- « `de` » (allemand),
- « `de-CH` » (allemand utilisé en Suisse)
- « `de-CH-1996` » (allemand utilisé en Suisse, orthographe de 1996)

Il existe deux types de filtrages : un filtrage de base et un filtrage étendu. Le filtrage de base exige des préfixes communs : « `de-CH` » correspond donc à « `de-CH` » ou « `de-CH-1996` », mais pas à « `de-Latn-CH` ».

Le filtrage étendu avec l'aide du joker « * » permet d'accepter toutes les valeurs d'une sous-étiquette « `de-*-CH` » correspond à « `de-CH` », « `de-CH-1996` », « `de-Latn-CH` » mais pas à « `de` » car il manque l'élément « `CH` ».

7.2. Consultation

Dans le cas de la consultation, on recherche le meilleur document parmi une liste de documents. L'utilisateur précise le choix de langue le plus précis possible, car il ne veut qu'un document en retour. « `de-CH` » peut rendre « `de` » ou « `de-CH` », mais pas « `de-CH-1996` ». Si aucune étiquette de langue ne correspond à la demande, la valeur par défaut est retournée.

Quelques applications de ce mécanisme pour trouver la meilleure version linguistique :

- La sélection d'un gabarit contenant le texte d'une réponse électronique automatique.
- La sélection d'un élément contenant du texte à inclure dans une page Web particulière.
- La sélection d'une chaîne de texte à inclure dans un journal des erreurs.
- La sélection d'un fichier son à jouer en invite d'un système téléphonique.
- La recherche par repli des informations « locales » d'un programme informatique : les messages à afficher, les conventions de tri, le calendrier à présenter dans la langue de l'utilisateur.

Lors du mécanisme de consultation, le choix de langues est tronqué progressivement à partir de la fin jusqu'à ce qu'une étiquette de langue correspondante soit trouvée dans la base des documents. Par exemple, en commençant avec le choix « `zh-Hant-HK` » (chinois, écriture traditionnelle, Hong Kong), la consultation recherche progressivement du contenu comme indiqué ci-dessous :

```
 Exemple d'un schéma de repli de consultation

Choix à réaliser : zh-Hant-HK
1. zh-Hant-HK
2. zh-Hant
3. zh
4. (défaut)
```

Le comportement de repli autorise de la flexibilité dans la recherche d'une correspondance. Sans repli, le contenu par défaut serait immédiatement retourné si un contenu correspondant exact n'était pas disponible. Grâce au repli, un résultat correspondant au mieux à la demande de l'utilisateur peut être fourni.

8 BCP 47

Pour éviter qu'une norme ne mentionne un RFC qui risque de devenir désuet, l'IETF recommande aujourd'hui aux auteurs de normes de faire référence à des *Best Current Practice* (« les Meilleures pratiques actuelles ») plutôt qu'à d'autres RFC. Les numéros de BCP ne changent pas, mais leur dernière version suit les derniers développements dans leur domaine.

Le BCP 47 regroupe les meilleures pratiques actuelles liées à l'indication de langue d'un document, d'une partie de document ou d'un objet. Le BCP 47 actuel est composé des RFC 5646 et 4647. La version précédente du BCP 47 comprenait les RFC 4646 et 4647 qui remplaçaient déjà le RFC 3066, désormais désuet, lequel avait déjà remplacé le RFC 1766.

9 Tout cela est bien beau, mais est-ce utilisé ?

Ces étiquettes de langues sont utilisées par de très nombreux produits, standards et normes parmi lesquels on peut nommer : XML, HTML, RSS, MIME, SOAP, SMTP, LDAP, CSS, XSL, CCXML, Java, C#, ASP, perl, Apache, IE, Firefox…

Certains processus comme le mécanisme qui sert à préciser la locale POSIX (la langue et les conventions locales) d'un processus utilisent encore le RFC 1766 (le bisaïeul du RFC 5646) :

```
LANG=fr_FR
setenv LC_COLLATE=de-DE@phone
```

Ce genre de syntaxe est très fréquent sur les machines Unix. La première ligne indique la langue à utiliser notamment pour l'affichage des messages que le système affichera. La deuxième précise que le tri devra se faire en considérant les données comme de l'allemand (« de ») d'Allemagne (« DE ») en utilisant la convention de tri du bottin téléphonique (« @phone »). Dans les deux cas, on remarque que les codes de langue et de pays correspondent respectivement à l'ISO 639 et l'ISO 3166 (mais dans leur version de 1988).

9.1 Recherche de la bonne version linguistique d'un page Web

Le protocole HTTP permet à un navigateur internet d'indiquer quelle langue l'utilisateur préfère lire à l'aide de l'entête `Accept-`

`language` qui accompagne chaque demande de document. La négociation de langue est très avantageuse pour naviguer sur les sites multilingues : le client obtient directement la version qui lui convient, sans perdre son temps à louvoyer parmi des pages qu'il comprend mal ou pas du tout, à la recherche d'un hyperlien vers la bonne langue. Le site <http://www.debian.org/> et celui du W3C (mais de manière très parcellaire[19]) l'utilisent.

L'`Accept-Language` de HTTP 1.1 permet de préciser un choix de langues codées sous la forme proposée par le RFC 1766. Toutefois, une révision de ce protocole (dénommé pour l'instant « HTTPbis ») permettra de définir celles-ci selon les RFC 5646 et 4647 (la dernière version du BCP 47 donc).

Comme nous l'avons vu, la recherche du bon document Internet à l'aide d'`Accept-Language` correspond à une consultation en termes du RFC 4647.

9.2 Recherche des ressources de programmes

Le même mécanisme de consultation et de repli pour trouver la ressource qui correspond le mieux au profil linguistique de l'utilisateur est utilisé par de nombreux environnements de développement comme Java ou C#.

Le programme écrit dans ces langages pourra ainsi accéder automatiquement aux valeurs qui correspondent le mieux à l'utilisateur. Un programme correctement conçu pourra ainsi aller chercher et afficher les messages, les libellés, les menus, les avertissements dans la langue qui correspond le mieux à l'utilisateur par le même processus que nous avons décrit dans la section *7.2 Consultation* ci-dessus.

9.3 Indiquer la langue, l'écriture, le pays dans HTML et CSS

On recommande d'ajouter, dans les documents HTML, un attribut `lang` à la balise html. On préfère le mettre sur la balise `html` plutôt que le body, car il existe au moins un passage de texte qui ne serait pas balisé en choisissant `body` : le titre (`title`). L'exemple ci-dessous déclare un document écrit en français de Belgique :

[19] La section des communiqués de presse du W3C utilise cette technique, voir par exemple <http://www.w3.org/2010/08/woff-pr.html>.

```
<html lang="fr-BE">
```

En XML, et donc en XTHML, on utilisera l'attribut `xml:lang` pour préciser la langue :

```
<html xml:lang="fr-BE" xmlns ="http://www.w3.org/1999/xhtml">
```

Ces attributs peuvent s'utiliser sur la quasi-totalité des balises où il est raisonnable de préciser une langue. La valeur de ces attributs correspond à un codet défini par le BCP 47. Parmi les balises où ces attributs sont fréquemment utilisés : `<p>` (le paragraphe) et `<span>` (le bout de texte). Par contre, on ne peut inclure d'attribut de langue sur `<br>` (passage à la ligne) ou `<hr>` (ligne horizontale, filet) qui n'ont pas de valeur linguistique.

Après avoir déclaré la langue globale du document, il est facile de mentionner qu'un passage est dans une autre langue par l'adjonction d'un attribut `lang`, `xml:lang` ou les deux sur la balise qui comprend le texte dans la langue étrangère :

```
<html lang="fr">
<body>
<p>Triple-patte s'exclama alors : <span lang="la">Timeo
Danaos et dona ferentes</span>. Homère aurait dit :
<span lang="grc">Φοβοῦ τοὺς Δαναοὺς καὶ δῶρα φέροντας</span>.</p>
```

Le langage de « stylage » CSS utilisé avec HTML connaît un type de sélecteur d'intérêt en matière linguistique : la pseudoclasse. La pseudoclasse est un sélecteur spécial qui définit une condition difficilement exprimable par les sélecteurs standard ou qui n'est pas liée à la structure du document. Une pseudoclasse nous intéresse particulièrement, il s'agit de `:lang(`*l*`)` où *l* représente une langue.

La pseudoclasse `:lang(l)` sélectionne des éléments d'une langue « l » sans que ces éléments aient nécessairement un attribut `lang`, il suffit que la langue courante soit celle mentionnée entre les parenthèses de `lang()`. Rappelons que les éléments HTML/XHTML héritent de l'attribut lang de leurs éléments supérieurs (« leurs parents »). Le stylage lié à la langue permet de régler avec précision la présentation d'une page ou d'un passage selon la langue en question. On peut de la sorte choisir une police particulière pour des passages particuliers : des polices touarègues pour mieux rendre l'original en tamachek, une belle police

coufique en arabe, etc. La recherche de correspondance est à nouveau une consultation au sens du RFC 4647, on cherche la meilleure correspondance avec comme stratégie de repli la troncature successive par la droite de l'attribut de langue (hérité au besoin) de l'élément que l'on doit « styler ».

C'est ainsi que dans le code ci-dessous :

```
<style>
:lang(ar) { color: blue; }
:lang(ar-MA) { color: red; }
</style>
<body>

<p>Il se tourna vers nous et dit, philosophe, <span lang="ar-
MA">inch Allah</span>.</p>

…
<p>Il se tourna vers nous et dit, philosophe, <span
lang="ar">inch Allah</span>.</p>

<p>Il se tourna vers nous et dit, philosophe, <span lang="ar-
ary">inch Allah</span>.</p>
```

Le premier « inch Allah » sera en rouge : la meilleure correspondance est celle pour `ar-MA`. Le deuxième sera en bleu, car « `ar` » est la meilleure valeur. Quant à la troisième invocation, elle s'affichera en bleu, car aucune pseudoclasse pour « `ar-ary` » n'a été définie, mais il en existe une pour « `ar` », qui correspond par troncature par la droite.

10. Conclusion

Indiquer la langue d'un texte ou d'un passage de ce texte s'avère utile pour de nombreux processus, qu'il s'agisse de la vérification orthographique, du repérage de texte comme Google, de trier ce texte ou encore d'en permettre la synthèse vocale.

La syntaxe des étiquettes linguistiques est définie par le BCP 47 qui est actuellement constitué des RFC 5646 et 4647. Chaque étiquette linguistique est composée d'une ou plusieurs « sous-étiquettes » séparées par des traits d'union. Si l'on excepte les étiquettes à usage privé et les étiquettes patrimoniales conservées pour des raisons de compatibilité arrière, les sous-étiquettes doivent se présenter dans l'ordre suivant :

- une sous-étiquette qui représente la langue
- une sous-étiquette optionnelle qui représente une langue plus précise (« extlang ») quand la première sous-étiquette fait référence à une macrolangue.
- une sous-étiquette optionnelle qui représente l'écriture
- une sous-étiquette optionnelle pour la région
- une série optionnelle de sous-étiquettes représentant les variantes
- une série optionnelle de sous-étiquettes représentant des extensions
- une série optionnelle de sous-étiquettes à usage privé.

Exemples :

es	représente l'espagnol
fr-CA	le français au Canada
yue-Hant-HK	le cantonais écrit en chinois traditionnel à Hong-Kong

À ce stade, il n'existe pas d'indicatif particulier pour représenter l'amazighe marocain commun enseigné dans les écoles marocaines et promu par l'IRCAM. C'est une lacune qu'il faudra sans doute combler dans les années à venir. La question a déjà été abordée dans plusieurs cercles d'experts, une des questions qui risque de se poser lors de la codification d'un nouvel indicatif pour l'amazighe sera de voir s'il faut le considérer comme une macrolangue (regroupant les langues tamazight, rifaine, etc.) ou comme une nouvelle langue supplantant ou regroupant des dialectes, un peu comme l'allemand standard le fit.

11. Remerciements

Nous tenons à vivement remercier l'IRCAM et plus particulièrement le directeur du CEISIC, Youssef Aït Ouguengay, pour leur accueil chaleureux et l'organisation du colloque international à Rabat au cours duquel cette communication a été présentée. Nous remercions également M. Ouguengay, son équipe, ainsi que Mme Aïcha Bouhjar directrice du CAL et son équipe, pour leur appui essentiel pour mettre à jour les normes de l'ISO et de l'IETF afin d'inclure les informations nécessaires pour

pouvoir étiqueter correctement les textes électroniques écrits en amazighe.

Bibliographie

Andries, P. (2008*). Unicode 5.0 en pratique*, Dunod éditions, Paris.

Bortzmeyer, S. (2009), *RFC 5646: Tags for Identifying Languages*, disponible à <http://www.bortzmeyer.org/5646.html>.

Phillips, A et Davis, M. (2009), *RFC 5646*, disponible à <http://www.rfc-editor.org/rfc/rfc5646.txt>.