Gestion des sites multilingues : comment utiliser la balise Hreflang ?

Retour au sommaire de la lettre

Domaine :	Recherche	**Référencement**
Niveau :	Pour tous	**Avancé**

Google propose depuis quelques mois la prise en compte de la balise Link Hreflang pour indiquer les différentes versions linguistiques d'une même page. Or, on s'aperçoit que cette balise est très souvent mal implémentée dans les sites web. Cet article a donc pour volonté d'expliquer comment Google détecte la langue d'un contenu ainsi que les différentes possibilités disponibles pour indiquer au moteur quelle est celle utilisée dans une page donnée et comment il faut les utiliser à bon escient...

Lors de leur exploration des pages du web, les moteurs de recherche rencontrent des pages rédigées dans les nombreuses langues différentes utilisées sur la Toile. L'identification exacte du langage employé est indispensable pour "classer" et "filtrer" correctement les pages par langue utilisée. Associer une langue à une page se révèle évidemment plus pratique pour les utilisateurs, mais aussi pour pouvoir appliquer les bonnes règles et les bons analyseurs lexicaux et syntaxiques aux textes à indexer.
Or la détection de la langue employée n'est pas du tout triviale pour un moteur de recherche... Nous allons d'abord voir pourquoi dans un premier temps, avant de nous intéresser à une solution proposée par les moteurs Google et Yandex uniquement pour "aider" ces moteurs à associer les pages à la bonne version linguistique : l'annotation <link rel="alternate" hreflang=[xxx]>. Nous détaillerons ensuite les cas d'utilisation de cette balise dont la manipulation est souvent mal comprise par les webmasters.

Le problème de la déctection de la langue sur les pages multilingues

Le premier écueil que rencontrent les moteurs pour identifier la langue d'une page est dans un premier temps la complexité des langues parlées sur Terre. Il existe tout un continuum de situations entre la langue officielle, les langues locales, les variantes régionales, les dialectes, les "patois locaux", les créoles, les niveaux de langage (ex : le langage SMS comparé au français littéraire), les usages (ex : l'arabe moderne / l'arabe classique).

Cette absence de critère linguistique permettant de séparer clairement langues et variantes, langues et dialectes, empêche de comptabiliser correctement le nombre de langues parlées sur Terre. Mais on parle de plusieurs milliers de langues différentes... Un moteur comme Google n'en gère que 130 environ.

Le second écueil est bien sûr que les sites sont parfois rédigés en plusieurs langues, et surtout, que plusieurs langues peuvent être présentes sur la même page !

La solution théorique : déclarer la langue du contenu dans le code de la page

La solution la plus simple pour permettre au moteur d'identifier la langue d'une page pourrait être d'analyser la langue déclarée par le webmaster. En effet, il est possible de déclarer, dans les en-têtes http: ou dans les en-têtes HTML la langue de la page.

Déclarer la langue utilisée est plus qu'une bonne pratique, c'est indispensable et il est clairement recommandé de le faire à chaque fois que c'est techniquement possible.

Notons au passage qu'il existe deux types d'annotation pour les langues, avec des objectifs différents :
- celles destinées à spécifier la langue primaire du document ;

- celles destinées à spécifier la langue de traitement du document.

La "**langue primaire**" est une méta-donnée qui s'applique à tout un document. Dans la pratique, c'est une indication qui s'adresse aux navigateurs web. Si une page est destinée à être affichée dans le navigateur d'un internaute réglé pour lire des pages en italien par défaut, on utilisera soit :

Content-Language: it (dans l'en-tête http:)

soit :

<meta http-equiv="Content-Language" content="it"> (dans l'en-tête html)

Notons que l'on peut déclarer plusieurs langues dans ce type de balises, en séparant les langues par une virgule. Exemple :

<meta http-equiv="Content-Language" content="it,da"> (italien + danois)

La langue de traitement est, quant à elle, une indication qui s'adresse à d'autres applications que le navigateur : les logiciels de traduction, ou les logiciels de correction grammaticale ou orthographique par exemple. Elle est spécifiée dans le code par l'attribut html "lang" ou "xml:lang". Ces annotations servent à indiquer la "vraie" langue utilisée dans la zone entourée par le conteneur. Cela signifie par conséquent qu'une seule valeur est possible par cet attribut.

- En HTML classique : <html lang="fr">

- En XHTML traité en tant que HTML : <html lang="fr" xml:lang="fr" ...>

- En XHTML traité en tant que XML (type de contenu application/xhtml+xml) : <html xml:lang="fr" ... >

Si une partie quelconque du contenu est rédigée dans une autre langue que celle déclarée dans la balise HTML, il suffit de l'indiquer dans les attributs lang et xml:lang de son élément conteneur. Par exemple, pour une citation en anglais dans un document en français :

<q lang="en">...</q>

Mais on ne peut pas faire confiance aux webmasters !

Dans la pratique, on constate que ces attributs, directives et annotations sont méconnus. Trop de webmasters n'utilisent pas ces possibilités de déclaration, ou pire, les utilisent mal, ce qui oblige les moteurs à chercher à reconnaître la langue en fonction du lexique utilisé dans une zone donnée de la page, et à ne pas faire confiance aux déclarations de langue !

La reconnaissance de la langue d'après le vocabulaire utilisé est relativement fiable : on pourra tester différents exemples avec l'outil en ligne de Translated : http://labs.translated.net/identificateur-langue/

Mais l'exercice présente ses limites lorsque les textes sont courts, ou rédigés en style télégraphique, en langage SMS, bourrés de fautes d'orthographes etc.

Par conséquent, les moteurs se trompent dans un certain nombre de cas, et vont donc prendre par exemple pour de l'anglais un bout de texte technique rédigé en français !

Le problème des variantes locales

Si on prend un cas "simple" comme l'anglais, un américain reconnaîtra assez vite qu'une page est rédigée en anglais du Royaume-Uni et non pour un internaute du Kansas. Les différences entre les textes rédigés dans les variantes locales sont en effet multiples :

- différences de graphie ("theatre", "colour" en UK, "theater", "color" aux US) ;
- différences de vocabulaire ("lorry", "bonnet" en UK, "truck", "hood" aux US).

Par contre, l'anglais américain est dominant sur les pages web, y compris en UK, donc ce texte tiré de l'aide de Google, rédigé en anglais américain (révélé par l'orthographe "organization" avec un z et non un s comme en anglais du Royaume Uni) :

Good account organization helps you make changes quickly, target your ads effectively, and, ultimately, reach more of your advertising goals.

ne peut pas être déclaré avec certitude comme rédigé en anglais américain pour des américains ! C'est encore plus vrai pour des pages de sites canadiens anglophones, qui mélangent allégrement - et de plus en plus - les habitudes britanniques ou américaines (en choisissant de plus en souvent la graphie américaine).

Les variantes entre français de Belgique ou français Canadien ne sont pas suffisamment marquées pour être reconnues.

Si l'on prend le cas du français, les différentes versions parlées au Canada, en France, en Suisse, en Belgique, sont trop proches pour être distinguées automatiquement avec un bon niveau de certitude. Le lexique est parfois différent, mais pas la graphie, ni la grammaire, et il faut donc un texte assez long pour identifier immédiatement l'origine géographique du rédacteur.

Par contre, un seul terme de vocabulaire reconnu par le lecteur suffira parfois à identifier cette origine (ex : "char" à la place de "voiture" ou "automobile" dans un texte).

Les quasi doublons dus aux différentes versions linguistiques : un problème épineux à résoudre

Le développement de versions multilingues des pages produit "de facto" des pages dont le contenu est très proche les unes des autres est un cas complexe. Ce phénomène est particulièrement net dans deux cas :
- lorsqu'on ne traduit que l'interface ;
- lorsqu'on "localise" subtilement les pages pour les adapter à des publics locaux.

1er cas : traduction de l'interface utilisateur uniquement

Si l'on prend par exemple un site d'annonces, le texte de l'annonce est rédigé dans la langue de l'offreur. Par contre, il est possible d'afficher l'interface dans une autre langue (comme le bouton "répondre à cette annonce"), sans que l'on traduise le contenu de l'annonce (c'est même la pratique la plus courante). Les pages affichant chacune une version de l'interface dans une langue différente ont donc un contenu majoritairement similaire : ce sont des quasi doublons.

2e cas : localisation de la page

Dans certains cas, la "localisation", c'est-à-dire l'adaptation des textes pour un public local, produit des quasi doublons encore plus spectaculaires.

Prenons un vendeur de piscines espagnol : ses descriptifs de piscine seront évidemment identiques dans sa version espagnole et dans la version mexicaine, sauf que "piscine" se dit "alberca" au Mexique, et "piscina" en Espagne.

Les deux pages seront donc identiques... à un mot près.

Parfois la localisation consiste à changer uniquement la monnaie et les prix, ce qui crée des pages également très similaires.

Quand le multilinguisme crée des doublons !

Dans les types de cas de quasi doublons qui viennent d'être signalés, les moteurs de recherche ont en général du mal à s'y retrouver sans qu'on les aide, et font trop souvent les mauvais choix. En général, ces pages sont réellement considérées comme des doublons et éliminées ou canonicalisées sauvagement dans le processus d'indexation (y compris en l'absence de balises link rel='canonical').

Il y a donc des chances pour que nos amis mexicains ne voient pas les pages de notre marchand de piscines espagnol, ou pire, que nos amis espagnols soient perplexes devant des "albercas" soit disant construites... en Espagne !

Le problème ici n'est donc pas uniquement un problème de détection de la langue : on cherche à obtenir aussi l'inverse d'une canonicalisation, c'est-à-dire que Google considère vraiment ses pages comme des pages différentes, et les indexe... au bon endroit !!

L'annotation <link rel="alternate" hreflang=...> à la rescousse

Google a donc décidé d'introduire une annotation supplémentaire pour permettre aux webmasters d'indiquer clairement dans quel index linguistique classer les pages, et pour indiquer également qu'un groupe de pages représente des variantes linguistiques de la même page. Rappelons ici que Google propose des index séparés :
- d'abord pour chaque pays ;
- ensuite par langue du pays.

Dans la pratique on cherche à cibler :
- soit une langue ;
- soit une langue ET une zone géographique.

La syntaxe de cette annotation est :

```
<link rel="alternate" hreflang="[code langue]" href="[url]" />
```

Remarque : cette balise est correctement supportée par Yandex et Google, mais pas par Bing.

IMPORTANT : Une erreur fréquemment observée est de ne placer qu'une balise link rel="alternate" dans le header pointant vers l'url par défaut ! Cette logique ressemble à celle du link rel="canonical", mais on cherche à faire l'inverse : faire indexer TOUTES les versions linguistiques, et si possible au bon endroit.

Il faut donc placer dans le header une balise par version linguistique, y compris celle présente sur la page en cours. Voici donc un extrait de l'entête d'une page de support de Google :

```
<link rel="canonical"
href="http://support.google.com/webmasters/bin/answer.py?hl=fr&answer=2620865" />

<link rel="alternate" hreflang="ar"
href="http://support.google.com/webmasters/bin/answer.py?hl=ar&answer=2620865">
<link rel="alternate" hreflang="bg"
href="http://support.google.com/webmasters/bin/answer.py?hl=bg&answer=2620865">
<link rel="alternate" hreflang="id"
href="http://support.google.com/webmasters/bin/answer.py?hl=id&answer=2620865">
<link rel="alternate" hreflang="ca"
href="http://support.google.com/webmasters/bin/answer.py?hl=ca&answer=2620865">
```

```
<link rel="alternate" hreflang="cs"
href="http://support.google.com/webmasters/bin/answer.py?hl=cs&answer=2620865">
<link rel="alternate" hreflang="sr"
href="http://support.google.com/webmasters/bin/answer.py?hl=sr&answer=2620865">
<link rel="alternate" hreflang="da"
href="http://support.google.com/webmasters/bin/answer.py?hl=da&answer=2620865">
<link rel="alternate" hreflang="de"
href="http://support.google.com/webmasters/bin/answer.py?hl=de&answer=2620865">
<link rel="alternate" hreflang="en"
href="http://support.google.com/webmasters/bin/answer.py?hl=en&answer=2620865">
<link rel="alternate" hreflang="es"
href="http://support.google.com/webmasters/bin/answer.py?hl=es&answer=2620865">
<link rel="alternate" hreflang="es-419"
href="http://support.google.com/webmasters/bin/answer.py?hl=es-
419&answer=2620865">
…
```

En réalité, il existe une trentaine de versions linguistiques donc la liste continue !

Comme pour la directive "Content-language", une syntaxe existe pour les en-têtes http:, ce qui permet de régler le cas des contenus de type pdf ou word par exemple.

```
Link: <[url]/>; rel="alternate"; hreflang="[code langue]"
```

Mais on peut aussi spécifier cette information dans le sitemap xml ! Cette possibilité permet de rendre l'implémentation de ces balises plus facile (pas besoin de toucher au code du site web avec cette implémentation). On trouvera la syntaxe pour les sitemaps ici : http://support.google.com/webmasters/bin/answer.py?hl=fr&answer=2620865

Quels codes utiliser pour l'attribut hreflang ?

Les codes à utiliser pour l'attribut 'hreflang' sont ceux définis par la norme ISO 639-1 (pour les codes de langue) et la norme 3166-1 pour les codes pays. L'ajout d'un code pays est optionnel :
- Si une page est destinée à tout le public anglophone, on mentionnera uniquement hreflang='en' ;
- Si une page anglophone est destinée au public américain, on mentionnera hreflang='en-us'.

Comment utiliser cette balise conjointement avec link rel="canonical"

Contrairement à ce qui a été longtemps expliqué dans l'aide de Google sur cette balise, utiliser la syntaxe hreflang conjointement avec des balises canonical **EST EN GENERAL DECONSEILLÉ** sauf exception!

Il faut bien comprendre l'impact de ces balises sur le comportement de Google. Lorsqu'elles sont utilisées conjointement :
- la balise link rel="canonical" va conduire à l'indexation d'une seule version (l'url canonique) ;
- la balise link rel="alternate" sert à afficher une url différente dans chaque version pays+langue.

Par conséquent, le résultat sera que l'internaute verra une seule version du snippet (celle de l'url canonique) et sera redirigé vers la bonne version pays. Certes, mais un visiteur espagnol ne sera pas forcément ravi de voir la version anglaise du snippet sortir dans les pages de Google !!

Conclusion : en règle générale, on utilisera la balise "hreflang" seule, sans balise canonical...

L'indication par défaut : x-default-hreflang

Récemment (en avril dernier : http://www.abondance.com/actualites/20130410-12453-google-ameliore-sa-prise-en-compte-des-balises-multilingues.html) Google a ajouté une nouvelle syntaxe pour indiquer qu'il existe une page d'atterrissage par défaut pour les langues qui ne sont pas supportées par un site web.

Si la page d'atterrissage est : example.com/defaultlanding, alors, il convient d'ajouter cette ligne dans les en-têtes de toutes les pages qui sont des variantes linguistiques de cette page (y compris la page d'atterrissage par défaut elle-même) :

```
<link rel="alternate" href="http://example.com/defaultlanding" hreflang="x-default" />
```

Conclusion : dans quels cas utiliser ces annotations hreflang ?

Contrairement à ce que certains webmasters ont pu croire (les forums de Google sont remplis de questions démontrant que ces balises sont mal comprises), les annotations hreflang ne remplacent pas les spécifications de la langue primaire, et des langues de traitement. Il est toujours primordial de fournir ces informations, même si les moteurs n'en tiennent pas toujours compte.

Les annotations hreflang servent uniquement à s'assurer que les différentes versions linguistiques d'une même page soient bien :
- toutes indexées, et non considérées comme des doublons ;
- et indexées dans la bonne combinaison index pays x index linguistique.

La portée de cette balise est donc limitée, mais particulièrement utile pour un site multilingue. Leur implémentation dans ce contexte est donc conseillée, mais le webmaster devra être particulièrement prudent dans son implémentation à ne pas générer d'effets de bords, notamment en cas d'utilisation conjointe avec une balise <link rel='canonical …>.

Par contre, si on ne veut pas que ces variantes soient indexées (dans les cas de traduction de l'interface), on n'utilisera pas l'annotation hreflang, mais une balise <link rel='canonical …> pointant, pour toutes les variantes, vers la page en version par défaut. Il est dommage que l'implémentation de ces balises et annotations, et leur utilisation, ne soit pas plus clairement expliquées dans l'aide de Google. Mais nous espérons que cet article vous aidera à y voir plus clair.

Liens et bibliographie

Outil d'identification de la langue d'un texte
http://labs.translated.net/identificateur-langue/

Les pages de support de Google sur la balise hreflang :
http://support.google.com/webmasters/bin/answer.py?hl=fr&answer=189077
http://support.google.com/webmasters/bin/answer.py?hl=fr&answer=2620865&topic=2370587&ctx=topic
http://googlewebmastercentral.blogspot.fr/2013/04/x-default-hreflang-for-international-pages.html

Infos sur la norme ISO 639-1 (codes langues) :
http://en.wikipedia.org/wiki/List_of_ISO_639-1_codes

Infos sur la norme 3166-1 (codes pays) :
http://en.wikipedia.org/wiki/ISO_3166-1_alpha-2

Philippe YONNET , ***Directeur de l'agence Search-Foresight / Groupe MyMedia. Président de l'association SEO Camp (http://www.seo-camp.org/)***