L'affichage BIDI est plein de surprises

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 14 juin 2010

http://www.bortzmeyer.org/affichage-bidi.html

L'algorithme BIDI d'Unicode est complexe et peut donner des résultats surprenants. Comment, lorsque BIDI a pris la mauvaise décision, forcer l'affichage que l'on veut ?

Certaines écritures sont de gauche à droite comme, aujourd'hui, l'alphabet latin et d'autres sont de droite à gauche, comme l'alphabet arabe. Lorsqu'un texte est entièrement dans une écriture de gauche à droite, ou entièrement de droite à gauche, pas de problème. C'est lorsqu'on mixe les deux types d'écriture que les problèmes surviennent. Prenons un exemple classique, le Hello world du BIDI, « Ari dit [Caractère Unicode non montré [1]][Caractère Unicode non montré][Caractère Unicode non montré] [Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré] et sourit. » Le texte en français, écrit en alphabet latin, est de gauche à droite puis le sens change subitement pour le texte en alphabet hébreu puis revient à son sens original pour le français.

Mixer les deux types d'écriture est courant dans le monde réel : regardez par exemple la page Web en arabe <http://www.ojuba.org/wiki/okasha/%D8%A7%D9%84%D8%B5%D9%81%D8%AD%D8%A9_%D8%A7%D9%84%D8%A3%D9%88%D9%84%D9%89> décrivant un module Apache, avec des sigles en alphabet latin que les programmeurs Web reconnaitront. Composer un tel texte n'est pas évident : l'algorithme standard pour cela est l'algorithme Bidi, normalisé dans l'"*Unicode Standard Annex #9*" <http://unicode.org/reports/tr9/>. Je ne vais même pas essayer de résumer ce texte complexe. Disons simplement que certains caractères, notamment les lettres, ont une **directionnalité** (de gauche à droite ou de droite à gauche). Le moteur de rendu doit suivre cette directionnalité. Que faire lorsque le caractère n'a pas de directionnalité (cas des chiffres ou de la ponctuation) ? La directionnalité utilisée est alors celle du dernier caractère ayant une directionnalité. Cela donne parfois des résultats surprenants. BIDI n'est qu'un algorithme, après tout, et ne comprend pas forcément toutes les subtilités. Prenons ce texte :

Voici mes photos de [Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré] (10 juin 2010).

1. Car trop difficile à faire afficher par LaTeX

Pourquoi ce résultat illisible ? Dans le texte ci-dessus, le moteur de rendu se met à écrire de droite à gauche lorsqu'il rencontre le premier caractère arabe et ne change qu'au premier caractère latin, le J de juin. Le 10 et la parenthèse ouvrante se retrouvent donc au mauvais endroit. Le texte erroné ci-dessus est pourtant conforme à l'algorithme BIDI et donc tous les navigateurs Web standard auront le problème. (Firefox, depuis au moins la version 52, ne respecte plus BIDI et donc affiche un résultat « correct ».)

Comment peut-on faire ? Il existe plusieurs solutions :

— On peut simplement changer le texte pour mettre des caractères ayant une directionnalité (des lettres, typiquement).
— On peut forcer la directionnalité en utilisant l'attribut HTML dir `<http://www.w3.org/TR/html401/struct/dirlang.html#h-8.2>`.
— On peut forcer la directionnalité avec les marques de directionnalité Unicode, les caractères U+200E et U+200F.

Voici un exemple de la première technique, le texte a été modifiée, BIDI a un comportement bien plus naturel :

Voici mes photos de [Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré] prises le 10 juin 2010. On n'a évidemment pas forcément envie de modifier son texte. Voici donc une seconde version, avec l'attribut `dir` de HTML (regardez le source de cette page pour voir comment c'est fait) :

Voici mes photos de [Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré] (10 juin 2010). Et la troisième solution, les marques Unicode ? L'"*Unicode Technical Report #20*" `<http://www.unicode.org/reports/tr20/>` déconseille leur usage pour HTML et XML. Mais, évidemment, tous les textes ne sont pas forcément dans ces formats. Dans le cas de documents en dehors de ce monde XML, ces marques sont une bonne solution. Voici un exemple de leur utilisation (U+200E est la marque gauche-à-droite et U+200F celle droite-à-gauche) :

Voici mes photos de [Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré][Caractère Unicode non montré] [Caractère Unicode non montré](10 juin 2010). Pour le cas particulier des noms de domaine, on peut consulter les règles BIDI du RFC 5893 [2].

À noter que le problème n'a évidemment rien à voir avec l'encodage du texte puisque HTML sépare complètement l'encodage `<http://www.w3.org/TR/charmod/>` du modèle de caractère (qui est toujours Unicode, quel que soit l'encodage du source).

2. Pour voir le RFC de numéro NNN, `https://www.ietf.org/rfc/rfcNNN.txt`, par exemple `https://www.ietf.org/rfc/rfc5893.txt`