Codage des lettres accentuées dans les fichiers sources Latex
Version 2.0

Par

Robert Rolland

http://www.acrypta.fr

Pour comprendre le problème de la saisie d'un texte en latex il faut considérer qu'en principe, les lettres accentuées et autres caractères non classiques (qui apparaissent sur les claviers français) sont saisis avec une commande « \ » donnée par le tableau suivant :

é	è	ê	ë	î	ï	à	â	ô	ù	û	ü	ç
\'e	\`e	\^e	\"e	\^\i	\"\i	\`a	\^a	\^o	\`u	\^u	\"u	\c c

§	£
\S	\pounds

Ceci n'est pas très commode, et on souhaite utiliser maintenant la saisie au clavier des lettres qui apparaissent sur le clavier, et en particulier les caractères accentués. Ceci met en jeu **deux mécanismes distincts** :

1) d'une part **le codage des caractères accentués** par l'éditeur de texte utilisé;
2) d'autre part l**e package à utiliser** pour que le compilateur latex comprenne le codage de ces caractères.

Je ne considèrerai ici que les deux codages suivants : **isolatin1** et **utf8**.

Si l'éditeur de textes code le source en isolatin1 alors dans le préambule du fichier latex on doit faire appel au package latin1 :

\usepackage[latin1]{inputenc}

Si l'éditeur de textes code le source en utf8 alors dans le préambule du fichier latex on doit faire appel au package utf8 :

\usepackage[utf8]{inputenc}

Jusque là pas de problème. Le problème surgit lorsqu'on change d'éditeur de texte (par exemple si on travaille sous linux avec « kile » qui par défaut code en utf8 (on peut modifier ce comportement) puis sous windows où les éditeurs standards codent en isolatin1. On se retrouve alors avec des éditeurs qui ne se comprennent pas (c'est un problème avec les éditeurs, pas avec latex).
Si on compose un document avec des textes provenant de diverses personnes utilisant des encodages différents alors là on a non seulement un problème d'éditeur de textes, mais aussi un problème latex pour la définition du package à utiliser.

On a donc intérêt à disposer de logiciels de conversion. Personnellement, j'ai choisi d'écrire deux logiciels : « utf82tex » et « iso2tex » qui transforment respectivement un fichier écrit avec un éditeur codant en utf8 (resp. en isolatin1) en écriture tex (avec les « \ »).

Voici l'utilisation de ces deux logiciels :
Considérons le texte suivant :

voici un texte accentué.

dont on a fait deux fichiers : un fichier « texteiso.txt » contenant le texte précédent codé en isolatin1, un fichier « texteutf8 » contenant le texte précédent codé en utf8. On peut transformer chacun de ces fichiers en un fichier contenant le texte tex pur :

voici un texte accentu\'e.

Les commandes sont les suivantes :

iso2tex texteiso.txt

utf82tex texteutf8.txt

Les fichiers obtenus sont respectivement :

texteiso.txt.tex

texteutf8.txt.tex

En conclusion, on peut travailler indifféremment avec un éditeur qui fait de l'isolatin1 ou de l'utf8. À la fin on traduit en texte tex pur (qui est de l'ASCII non étendu pur) et donc qui est compréhensible par tout éditeur de texte. En outre une fois sous cette forme, le compilateur latex n'a plus besoin ni du package « latin1 » ni du package « utf8 » (les instructions \usepackage[latin1] {inputenc} ou \usepackage[utf8]{inputenc} sont inutiles).

Bien évidemment, aussi bien sous les bureaux de Linux que sous celui de windows, il est possible de se passer de la ligne de commande en créant une icône pour le programme et en glissant l'icone du fichier à transformer sur l'icone du programme.

Le package contient aussi les deux programmes suivants :

iso2utf8
utf82iso

qui permettent de passer du codage isolatin1 au codage utf8 des caractères. Il existe des commandes qui réalisent ce travail (iconv sous linux).

Les sources C de tous les programmes sont fournis dans le package.

Annexe : les codes

Dans les programmes précédents je n'ai traduit que les caractères suivants (ce qui semble suffisant pour l'application en vue) :
é è ê ë î ï à â ô ù û ü ç § £

Si on n'a pas les tables de codage sous la main on peut retrouver facilement les codes (aussi bien utf8 que isolatin1) de la manière suivante :

On crée deux fichiers contenant le texte précédent (é è ê ë î ï à â ô ù û ü ç§ £), l'un avec un éditeur qui code en utf8 et qu'on appellera « testutf8.txt », l'autre avec un éditeur qui code en isolatin1 et qu'on appellera « testiso.txt ».
On lit alors avec la commande « od » (sous linux) les suites d'octets (en hexadécimal) de ces deux fichiers :

od -A d -t x testutf8.txt

0000000 c320a9c3 aac320a8 20abc320 c320aec3
0000016 a0c320af 20a2c320 c320b4c3 bbc320b9
0000032 20bcc320 000aa7c3

od -A d -t x testiso.txt

0000000 20e820e9 20eb20ea 20ef20ee 20e220e0
0000016 20f920f4 20fc20fb 00000ae7

On obtient alors par exemple pour le codage utf8 (attention chaque bloc se lit de droite à gauche) :

c3 a9 20 c3 a8 20 c3 aa 20 c3 ab 20 ...

ce qui montre que « é » se code sur 2 octets : « c3 a9 », etc ... (le codage 20 est celui de l'espace blanc, le 0a final (suivi de 00) est le « aller à la ligne»

Ce même « é » est codé sur un octet « e9 » en isolatin1.