codepage un style pour traiter différents code de page dans le même document

Alain Aubord

le 29 novembre 2008

Résumé

Cet article décrit les nouvelles commandes disponibles pour traiter un code de page[1] ainsi que les problèmes rencontrés pour réaliser cette implantation uniquement en TEX.

1 Introduction

La version 3 de TEX permet de composer plus facilement des documents dans une autre langue que l'anglais. En effet, cette version autorise l'usage de caractères codés sur 8 bits (valeurs de 0 à 255) dans le texte source et de multiples tables de césures dans un format.
Ces améliorations étaient nécessaires, mais pas suffisantes. En effet, il est encore nécessaire de disposer de styles adaptés aux différentes langues ainsi que des polices de caractères contenant des signes spécifiques (comme les caractères accentués[2], les guillemets français...).
Le style Babel et les polices Dc ont été développés pour résoudre ces problèmes. Cependant, il reste encore un problème de taille: les différents codes de page existant sont incompatibles. Un document écrit dans un code de page doit être converti dans un nouveau code de page lorsqu'il est utilisé sur un ordinateur avec un système d'exploitation différent ou avec des polices de caractères dont l'encodage correspond à un autre code de page (comme les polices Dc). La gestion des différents codes de page peut se faire de plusieurs manières en TEX:

- Par l'utilisation d'un programme externe de conversion de caractères. Lorsqu'un tel programme est bien intégré dans TEX (comme les tables Tcp définies par EmTeX), cette solution est très efficace et agréable à utiliser. Son principal désavantage est le manque de portabilité.

- Par l'utilisation de polices virtuelles[3]. Cette solution est très bien intégrée dans TEX[4]. Cependant, il est nécessaire de fournir les fichiers décrivant la métrique et le ré-encodage de chaque police virtuelle utilisée dans un document.

- Par l'utilisation de la composition pour obtenir un caractère accentué (i.e.`\'e` pour é). Cette solution fonctionne remarquablement bien avec les polices Dc. Le style fourni pour l'utilisation de ces polices redéfinit très habilement les commandes qui dessinent un caractère accentué. Chaque fois qu'un caractère accentué existe dans la police, il est employé (en lieu et place de la composition). Ce système est donc parfaitement intégré dans TEX, mais il est très désagréable de taper et de relire les caractères accentués suivant cette méthode.

- Par l'utilisation du paquet codepage qui est totalement compatible avec TEX (version 3). Ce paquet offre en outre la possibilité de composer avec TEX un document prévu pour un autre

[1] Un code de page est une convention qui définit une association unique entre un caractère et un nombre qui le représente dans un ordinateur.

[2] La composition, c'est-à-dire l'utilisation d'une commande de positionnement d'un accent suivie par la lettre à accentuer, présente l'avantage de pouvoir accentuer n'importe quel caractère mais empêche la césure correcte d'un mot accentué.

[3] Une police virtuelle est un mécanisme fourni avec TEX version 3 qui permet de changer le vecteur d'encodage d'une police de caractères, de composer une nouvelle police de caractères à partir de plusieurs polices etc.

[4] Certains vieux pilotes de périphériques ne savent pas utiliser les polices virtuelles.

code de page que celui de son propre ordinateur et de déclarer des exceptions de césures contenant des syllabes accentuées.

Ce paquet transforme tous les caractères supérieurs à 127 en caractère «actif» (chaque caractère actif est une commande TEX) et il nécessite la transmission de quelques fichiers annexes pour pouvoir composer un document sur un autre système.

Comme on le voit, il n'y a pas de solution parfaite dans tous les cas. Le problème provient d'un manque de norme dans la manière de coder les symboles spéciaux[5].

1.1 Le problème des césures

TEX effectue une césure automatique des mots en se basant sur une table décrivant quelles syllabes peuvent être coupées. Cette table est inclue dans le format[6] et sa forme est figée tant qu'un nouveau format n'est pas reconstruit. Lors de la construction d'une table de césure toutes les syllabes sont converties en minuscules (en utilisant le `\lccode` de chaque caractère) avant d'être enregistrées. Pour trouver une césure lors de la composition d'un document, le texte à couper est traduit en minuscule et comparé avec les tables pré-définies.
Lorsque les tables de césures utilisées lors de la construction du format et le texte du document utilisent des codes de pages différents, les caractères (dont la valeur est plus grande que 127) du texte à composer et ceux utilisés pour les tables de césure ne correspondront jamais. Aucune césure ne pourra alors être trouvée (des exceptions existent cependant).
Il est évidemment possible de modifier la table de césure et de regénérer un nouveau format, cependant cette solution n'est pas toujours possible ni forcément souhaitable pour plusieurs raisons:

- le fichier qui décrit la table de césure devrait être identique quel que soit le système pour obtenir une césure identique partout[7].
- la génération d'un format est une opération complexe et délicate qui peut exiger la possession d'autorisations spécifiques (sur une machine partagée entre plusieurs utilisateurs).

La solution proposée par le paquet `codepage` pour résoudre ce problème consiste en une nouvelle commande `\MakeHyphenationLetter` qui permet de modifier les valeurs associées à un caractère codé au-delà de 127 pour utiliser des valeurs compatibles avec la commande `\hyphenation`.

2 L'interface utilisateur

Cet interface a été conçu pour être «le plus simple possible». Seules deux macros (commandes) et quelques constantes sont définies.
Les différents codes de page possibles sont définis par des constantes:

`\FourThreeSeven` pour le code de page 437 du PC; ce code de page est surtout utilisé avec les systèmes ayant l'anglais comme langue principale;

`\EightFiveZero` pour le code de page 850 du PC. Ce code de page est semblable au code 437 mais il contient moins de signes semi-graphiques et mathématiques et plus de lettres (comme les caractères majuscules accentués). Ce code de page est surtout utilisé avec d'autres langues que l'anglais.

[5] Si un standard unique devait voir le jour, il serait défini sur 16 bits (valeur de 0 à 65535).

[6] Un format est un ensemble de commandes pré-définies qui sont analysées et enregistrées sous une forme spéciale par une variante du programme TEX appelée IniTEX. Le programme TEX peut ensuite charger en mémoire très rapidement un format.

[7] Un exemple très concret de cette situation est l'utilisation de la table de césure anglaise pour des textes français. Les tables de césures anglaises sont disponibles et identiques quelle que soit l'implantation de TEX. Lorsqu'on veut transmettre un document écrit en français à un correspondant dont on n'est pas certain qu'il puisse disposer d'une table de césure française, il est toujours possible d'utiliser le table de césure anglaise et d'indiquer les exceptions de césure avec la commande `\hyphenation`. Cette méthode fonctionne relativement bien si l'on prend la précaution de redéfinir les commandes qui dessinent les accents sur les caractères (comme cela est fait dans le paquet `codepage`).

`\IsoOne` ISO Latin set 1. Cette norme est utilisée par de nombreux systèmes (la plupart des systèmes UNIX et WINDOWS). L'encodage Dc des polices TEX suit cette norme pour certains caractères (en particulier toutes les lettres). Lorsque ce code de page est utilisé les caractères qui correspondent exactement aux caractères des polices Dc ne deviennent pas «actifs» mais gardent le «catcode» de lettres. Les caractères se dessinant en utilisant le mode mathématique de TEX restent cependant «actifs».

`\Mac` pour les caractères du MAC INTOSH.

Le choix de l'encodage des polices de TEX est défini par deux autres constantes:

`\CM` définit l'encodage standard de TEX. Ce code est défini sur 128 positions, les caractères accentués sont dessinés par composition.

`\DC` pour le nouvel encodage défini pour les polices TEX. L'utilisation de ce code implique l'usage des polices Dc. Lorsque ce code est choisi, les caractères existant dans un code de page sont simplement convertis en caractère équivalent des polices Dc.

Les symboles mathématiques sont dessinés en utilisant les commandes et le mode mathématiques de TEX, les autres symboles sont en général ignorés.
Les deux commandes principales sont:

`\codepage#1#2` cette macro accepte deux paramètres: le code de page du document, et le type d'encodage choisi pour les polices TEX.

Après l'appel à la macro `\codepage` tous les caractères supérieurs à 127 sont «actifs» (ils se comportent comme des commandes).

`\MakeHyphenationLetter#1#2` cette macro n'existe que si le code de page Dc est choisi. Elle sert à modifier les valeurs associées à un caractère supérieur à 127 pour qu'il puisse être utilisé avec la commande `\hyphenation`. Pour que ces modifications restent locales, `\MakeHyphenationLetter` doit donc toujours être appelée dans un groupe (entre des accolades `{` et `}`). Pour des raisons techniques ces modifications ne sont pas univoques, il est possible que plusieurs caractères aient des valeurs identiques après transformation. Cette situation n'est pas trop grave car elle est peu fréquente et n'entrave pas la césure (du moins dans le cas de la langue française), un message signalant un conflit potentiel est tout de fois émis.

`\CurrentEncoding` Cette commande contient la valeur du système d'encodage choisi. Sa valeur n'est définie qu'après l'appel à la commande `\codepage`.

Toutes les commandes qui servent à dessiner des accents sur des caractères sont redéfinies. Les valeurs originales des commandes sont sauvées avec un nouveau nom:

Nouveau nom	**Commande Originale**
`\Grave`	`\‘`
`\Circumflex`	`\^`
`\Tilda`	`\~`
`\OverBar`	`\=`
`\UnderBar`	`\b`
`\Join`	`\t`
`\HungarUmlaut`	`\H`
`\Acute`	`\’`
`\Diaresis`	`\"`
`\Breve`	`\u`
`\OverDot`	`\.`
`\UnderDot`	`\d`
`\Tcheche`	`\v`
`\Cedille`	`\c`

Voici encore quelques commandes définies par le paquet. Ces commandes devraient être d'un usage exceptionnel:

`\TRtrue` enclenche le mécanisme qui transforme chaque caractère supérieur à 127 pour qu'il soit imprimé correctement par TEX. C'est la valeur par défaut !

`\TRfalse` produit l'effet inverse de la commande précédente. Cette commande est surtout utile lorsque des caractères doivent être écrits dans des fichiers auxiliaires (comme les fichiers servant à construire les index ou les glossaires). Il est alors souvent plus judicieux de convertir les caractères spéciaux lors de la relecture de ces fichiers plutôt qu'à leur écriture.

`\AllActive` transforme tous les caractères supérieurs à 127 en caractère actifs.

`\AllOther` transforme tous les caractères supérieurs à 127 en caractère sans signification particulière pour TeX (leur `\catcode` vaut 12).

`\og` dessine les guillemets français ouvrants (définis seulement lorsque les polices CM sont choisies).

`\fg` dessine les guillemets français fermants (définis seulement lorsque les polices CM sont choisies).

`\atcatcode` qui est une constante qui contient la valeur du `\catcode` du caractère @ lorsque cette valeur est modifiée par le paquet CODEPAGE. Autrement, elle est identique à la commande `\relax`.

3 Un bref exemple (pour LaTeX2e)

```
\documentclass{article}
   \usepackage{t1enc,codepage}
   {\MakeHyphenationLetter{\EightFiveZero}{è}
    \hyphenation{sys-tè-me sys-tè-mes}
   }
   \codepage{\EightFiveZero}{\DC}
\begin{document}
   Unix est un joli système d'exploitation.
\end{document}
```

Cet exemple fonctionnera correctement si tous les caractères au-delà de 127 sont disponibles. Certaines implantations de TEX (comme EMTEX) transforment tous les caractères au-delà de 127 en caractère 127 ou utilise des tables spéciales (les tables TCP pour EMTEX).
Pour obtenir un comportement correct avec EMTEX, il faut utiliser l'option `/8` lorsque le format est généré.

Références

[1] M. Goosens, F. Mittelbach, and A. Samarin *The LaTeX Companion*, 1994, Addison-Wesley.

[2] V. Eijkhout *TEX by Topic A TEXnician's Reference*, 1992 Addison-Wesley

[3] R. Seroul *Le petit livre de TEX*, 1989, InterÉdition