Initiation au codage des fichiers & TeXShop*

v0.3.0–2015/12/27

H. Schulz & R. Koch

1 Introduction

Les utilisateurs de TeXShop sont souvent confrontés, lors de l'ouverture ou de la composition de fichiers, au fait que le texte affiché dans le source ou le document composé ne corresponde pas à ce qui devrait figurer : les caractères sont en désordre ou incorrectes. C'est généralement un problème de *codage* — l'éditeur ou TeX, ou les deux à la fois, n'interprètent pas correctement les données saisies. Ce document n'est qu'une initiation au codage. Il n'est certainement *pas* exhaustif ; il ne traite que des codages les plus fréquents à ce jour.

2 Définition du codage de fichier

Alors que nous pourrions généralement croire qu'un fichier source .tex soit composé de caractères, en réalité, comme tous les fichiers informatiques, il n'est juste qu'une longue série de nombres entiers, compris entre 0 et 255. Les informaticiens appellent ces nombres entiers *octets*.

Toutes les autres données informatiques doivent, d'une façon ou d'une autre, être codées en octets. Le codage en octets le plus courant pour un texte ordinaire est appelé ASCII ; il code tous les caractères présents sur une machine à écrire américaine ordinaire. Ainsi, les caractères de A à Z sont codés de 65 à 90, les caractères a à z vont de 98 à 123. Le caractère espace est codé par l'octet 32, et les chiffres, les parenthèses et les caractères de ponctuation sont codés par d'autres octets.

À l'origine, TeX devait être saisi en ASCII. Ce qui était suffisant aux États-Unis, c'est révélé fastidieux en Europe occidentale, où les accents, trémas, points d'interrogation culbutés, et autres caractères diacritiques sont utilisés ; des macros ont été nécessaires pour produire ces caractères, ce qui empêcha les coupures de mots. Des problèmes encore plus compliqués ont surgi quand TeX a été utilisé dans le Proche et l'Extrême-Orient.

Le codage ASCII utilise seulement les octets 0 à 127. Il est donc possible d'utiliser les octets 128 à 255 pour coder d'autres caractères. Actuellement, de nombreux codages différents utilisent ces octets pour afficher des caractères supplémentaires.

3 Extension de la table de caractères

Les trois codages étendus les plus utilisés sur Mac sont MacOSRoman, ISOLatin1 et IsoLatin9 [1].

Le codage MacOSRoman est un vestige d'une époque antérieure à OS X et, sans surprise, propre au Mac. Son utilisation n'est plus encouragée.

Le codage IsoLatin1 enrichi le codage ASCII avec les caractères accentués utilisés dans les langues d'Europe occidentale.

IsoLatin9 ajoute le symbole de l'euro, €, au codage IsoLatin1 ainsi que quelques autres changements.

*Traduit par René Fritz le 9 janvier 2016.

1. Dans ce document, nous emploierons la notation utilisée dans la directive de codage de TeXShop. Voir le tableau dans la section (8) à la page 5.

3.1 Autres codages utilisés avec TEX

Les autres codages sont, IsoLatin2 pour les langues d'Europe centrale, IsoLatin5 pour le turque et Iso8859-7 pour le grec. Plusieurs codages sont disponibles pour le russe et les langues cyrilliques. D'autres codages sont disponibles pour le coréen et le chinois ; mais les langues d'Extrême-Orient utilisent bien plus que 256 symboles, de sorte que ces codages ne sont pas très satisfaisants.

3.2 Et Windows

Windows Latin 1 est une version d'IsoLatin1 dont certains caractères sont codés en différents endroits, tels que définis par Microsoft Corp. Vous pouvez utiliser ce codage lorsque vous recevez des fichiers de gens qui emploient Windows.

3.3 Défaut majeur

Les différents codages ont été développés indépendamment par des sociétés informatiques étant donné que leurs produits sont vendus dans un nombre croissant de pays.

Malheureusement, les fichiers texte ne disposent pas d'un préambule indiquant le codage utilisé pour les générer. Ainsi, TEXShop ne dispose d'aucun moyen pour ajuster automatiquement le codage au divers fichiers lors de leur l'arrivée. Certains éditeurs de texte ont intégré des heuristiques pour essayer de deviner le bon codage, mais TEXShop ne peut pas les utiliser, car ceux-ci ne travaillent qu'à 90 % du temps, et une conjecture erronée peut tout chambouler.

4 Unicode

Comme le marché de l'informatique se développe partout dans le monde, les entreprises informatiques ont fini par se rendre à la raison et ont créé un consortium pour développer une norme, appelé *Unicode*, qui englobe tout. Le but d'Unicode est de coder tous les symboles couramment utilisés à travers le monde, y compris romain, grec, cyrillique, arabe, hébreu, chinois, japonais, coréen… Unicode peut même coder les hiéroglyphes égyptiens et récemment a pris en charge les symboles mathématiques.

Tous les systèmes informatiques modernes, y compris Macintosh, Windows, Linux et Unix, prennent désormais en charge Unicode. En interne, TEXShop et d'autres éditeurs Macintosh décrivent les caractères en utilisant Unicode et peuvent accepter un texte qui combine plusieurs langues : latin, grec, cyrillique, arabe, chinois… TEXShop peut même comprendre si l'arabe, l'hébreu et le persan sont écrits de droite à gauche. Pour saisir ces diverses langues, il faut activer les claviers dédiés en allant dans Préférences système… → Langue et région → Préférences Clavier… → Méthodes de saisie. Ces panneaux ont changé dans les versions récentes d'OS X ; dans El Capitan, sélectionnez un clavier sur la gauche, ou cliquez sur « + » en-dessous de la liste pour afficher les langues supplémentaires et ajouter leur clavier respectif.

Parce qu'il y a bien plus que 256 symboles, Unicode décrit des symboles en utilisant des entiers beaucoup plus longs. Unicode proscrit la structure interne de ces nombres, mais définit plusieurs façons différentes d'écrire le texte sur le disque. Le codage Unicode le plus populaire est UTF-8, mais UTF-16 et d'autres encore sont disponibles.

Le grand avantage d'UTF-8 est que les caractères ASCII ordinaires conservent leur forme d'un seul octet dans le fichier codé. Ainsi, les fichiers ASCII ordinaires restent valables en UTF-8.

Avec la plupart des codages en octets comme IsoLatin1, IsoLatin9, etc., chaque séquence d'octets constitue un fichier légal. Si vous ouvrez un tel fichier avec un mauvais codage, le fichier apparaît comme d'habitude, mais certains symboles seront incorrectes. Si quelqu'un en Allemagne, qui utilise IsoLatin9, collabore avec quelqu'un aux États-Unis qui emploie MacOSRoman, et que leur document est rédigé en anglais, ils peuvent ne pas remarquer cette disparité jusqu'à ce qu'ils relisent les références et découvrent que les accents et trémas ont disparus.

Cependant, toutes les séquences d'octets ne constituent pas des fichiers légaux en UTF-8, car les symboles non-ASCII sont convertis en octets en utilisant un code un peu compliqué. Dans l'exemple précédent, si le collaborateur allemand utilise IsoLatin9 et le collaborateur américain utilise UTF-8 et que le collaborateur allemand emploie dans les références des symboles non-ASCII comme les trémas, alors TEXShop rendra compte d'une erreur quand il essaiera d'ouvrir le fichier IsoLatin9 en UTF-8. TEXShop affichera alors un message d'erreur et offrira d'ouvrir le fichier dans le codage par défaut, dans ce cas IsoLatin9. Les nouveaux utilisateurs trouvent ces messages d'erreur bien plus déroutant que ceux obtenus pour des symboles parfois incorrects, et c'est une des raisons pour laquelle le codage par défaut de TEXShop n'est habituellement pas UTF-8.

Si d'autre part, les deux auteurs ont mis UTF-8 Unicode comme codage par défaut. Ce codage conserve tout ce qui est tapé dans TEXShop, donc aucun caractères curieux n'est perdu. L'HTML et d'autres codes sont généralement enregistrés en UTF-8, si bien que TEXShop peut être utilisé comme un éditeur de texte plus général. Par ailleurs, si un fichier TEX d'une source extérieure n'est pas en UTF-8, nous obtenons un *warning*. L'astuce est alors de laisser TEXShop ouvrir le fichier en IsoLatin9 et d'examiner la ligne inputenc du fichier qui vous indiquera le codage qui a été effectivement utilisé. Puis de fermer le fichier *sans apporter de modifications* et de l'ouvrir à l'aide du dialogue Ouvrir... en choisisant manuellement le bon codage. Une fois que le fichier est ouvert avec le bon codage vous pouvez ajouter la ligne directive de codage de TEXShop, pour ce codage, et l'enregistrer pour une utilisation future.

Toutes les méthodes de codage discutés ici, y compris Unicode, ignorent l'italique, le soulignement, le corps, la police, la couleur, etc. Elles codent juste les caractères. Il appartient aux utilisateurs de spécifier des attributs supplémentaires d'une autre manière. Par exemple, lorsque le programme d'Apple TextEdit est utilisé en mode *texte*, un utilisateur peut changer la police ou sa taille dans le document entier, mais pas pour chacune de ces sections séparément. Si le document est enregistré sur le disque et puis rechargé, les changements de police seront perdus. D'autre part, un traitement de texte comme Microsoft Word ou Pages d'Apple, contrôle d'avantage les polices, leur taille et autres... Ces programmes produisent du texte avec un codage propriétaire qui n'est lisible que par eux-mêmes. Mais le fichier garde toutes les informations attribuées.

Alors que tous les ordinateurs modernes prennent en charge Unicode, leurs jeux de polices renferment des symboles pour seulement une petite partie du monde Unicode. Beaucoup de polices possèdent un caractère spécial, souvent une boîte, pour signaler l'absence d'un caractère. Ainsi, si vous voulez écrire en arabe ou en hébreu, vous devez choisir une police qui contient leurs symboles. Les ordinateurs modernes prennent en charge une grande variété de symboles parce que l'industrie informatique couvre le monde, mais les symboles Unicode peu courants ne peuvent pas être couverts par une unique police prévue à cet effet.

5 Dualité de vue : TEXShop et TEX

Lorsque l'utilisateur sélectionne un codage approprié, il doit configurer à la fois TEXShop et le bon moteur TEX pour utiliser ce codage. Ces deux tâches posent différents ensembles de problèmes.

Aux États-Unis et dans d'autres pays anglophones, les utilisateurs peuvent souvent ignorer totalement les codages. Le codage par défaut de TEXShop supporte l'ASCII ; et TEX et LATEX prennent en charge l'ASCII depuis toujours. Donc, il n'y a rien à faire.

En Europe occidentale, les utilisateurs doivent prendre un peu plus de soins. Couramment, le codage par défaut de TEXShop en IsoLatin9, sera suffisant pour leurs besoins. Mais ils doivent configurer TEX et LATEX comme décrit ci-dessous, et choisir avec précaution des polices qui renferment les accents, trémas et autres... Les étapes nécessaires sont faciles à réaliser.

En Russie et en Europe de l'Est, les utilisateurs doivent prendre des mesures semblables, mais les auteurs de ce document ne sont pas bien informé sur les bonnes configurations, ils vous suggèrent d'obtenir de l'aide auprès d'amis qui utilisent déjà TEX.

Les utilisateurs de l'Extrême-Orient et du Moyen-Orient, et les chercheurs travaillant sur des projets multi-langues, auront besoin de consulter d'autres sources pour les configurations détaillées. Ces utilisateurs devraient certainement se tourner vers XeTeX et LuaTeX, parce que ces extensions de TeX utilisent directement Unicode et sont bien plus capables de gérer les langues pour lesquelles Unicode devient essentiel. XeTeX et LuaTeX peuvent, tous les deux, composer presque tous les fichiers sources TeX et LaTeX standard, mais ont des codes supplémentaires pour prendre en charge Unicode. Le gros problème avec ces langues est que seules les polices qui les supportent doivent être choisies. Pour simplifier ce problème, à la fois, XeTeX et LuaTeX permettent aux utilisateurs d'employer toutes les polices du système installées sur l'ordinateur.

6 Indication du codage utilisé pour charger et sauvegarder les fichiers à TeXShop

Pour définir le codage par défaut de TeXShop, il faut ouvrir les Préférences... de TeXShop. Sélectionner l'onglet Document. Dans la deuxième colonne, trouver la section Encodage. Sélectionner le codage souhaité dans le menu déroulant. Sélectionner ISO Latin 9 pour obtenir le codage courant par défaut, utile dans les pays anglophones et d'Europe occidentale. Vous devez sélectionner UTF-8 Unicode ou UTF-16 Unicode si vous voulez préserver tout ce qui est tapé dans l'éditeur TeXShop. Si vous prenez n'importe quel autre codage, certains caractères que vous aurez tapés dans TeXShop seront perdus si vous enregistrez et rechargez ensuite. D'autre part, UTF-8 ne fonctionne pas bien avec certaines extensions LaTeX, comme cela est expliqué plus loin.

Pour définir le codage d'un fichier, TeXShop dispose d'un mécanisme propre, indépendant du choix par défaut de l'utilisateur ou des choix des fenêtres de chargement et d'enregistrement. Pour faire en sorte que le codage utilisé pour lire ou écrire un fichier particulier soit UTF-8, vous devez ajouter la ligne suivante dans les vingt premières lignes du début du fichier :

```
% !TEX encoding = UTF-8 Unicode
```

Ceci se réalise aisément en sélectionnant dans les Macros la commande Encoding. Une boîte de dialogue apparaîtra où vous pourrez sélectionner le codage approprié. À le fermeture du dialogue, la ligne sera placée en haut du fichier et remplacera toute ligne de codage existante.

Si une telle ligne existe, le codage indiqué sera utilisé, remplaçant tous les autres paramétrages du codage, *sauf* si la touche option *(alt)* est enfoncée pendant toute l'opération de chargement ou de sauvegarde.

Beaucoup d'utilisateurs en Europe occidentale préfèrent utiliser IsoLatin9 comme codage par défaut afin de pouvoir lire facilement les fichiers des collaborateurs, mais placent en tête de leurs fichiers modèles la ligne qui fixe le codage en UTF-8, pour que leurs propres fichiers soient codés en UTF-8.

Il est également possible de définir le codage utilisé pour lire un fichier en ouvrant le fichier explicitement depuis TeXShop. La boîte de dialogue résultante possède dans sa partie inférieure un menu déroulant qui permet de choisir le codage à utiliser pour ce fichier en particulier [2] (notez que la ligne « %!TEX encoding = » écrase cette commande).

L'enregistrement explicite d'un fichier depuis TeXShop produit un dialogue d'enregistrement qui possède un menu déroulant semblable permettant de définir le codage.

REMARQUE. — vous ne pouvez pas facilement changer le codage d'un fichier. La meilleure chose à faire est de copier l'ensemble du document en un nouveau et de sauver ce dernier avec le bon codage. L'utilisation de la directive de codage de TeXShop avant d'enregistrer le nouveau fichier pour la première fois est sans aucun doute recommandée.

2. Sous El Capitan vous devez d'abord appuyer sur le bouton Options pour obtenir ce menu déroulant.

7 Indication du codage du fichier à LaTeX

Votre moteur de composition doit *connaître* le codage utilisé pour enregistrer chaque fichier source afin que le source d'entrée et les glyphes produits soient synchronisés. Pour LaTeX, cela se fait généralement en incluant la commande suivante dans le préambule du source :

```
\usepackage[latin9]{inputenc}
```

Les valeurs habituelles pour les autres codages sont données dans le petit tableau, à la fin de ce document.

Cette ligne n'est pas nécessaire lorsque le source est communément codé en ASCII.

utf8 constitue une des valeurs légales de codage avec inputenc. Cette ligne fonctionne en Europe occidentale, mais pas dans les situations nécessitant une utilisation avancée d'Unicode. En cas de doute, il est utile de lire la documentation sur inputenc. Pour cela, aller dans le menu Aide de TeXShop et sélectionner Afficher l'aide pour le package..., et remplir le nom de l'extension demandée avec inputenc.

En Europe occidentale, les utilisateurs entrent habituellement *quatre* commandes dans le préambule. Voici ces quatre lignes pour les Allemands.

```
\usepackage[german]{babel}
\usepackage[lmodern]
\usepackage[T1]{fontenc}
\usepackage[latin9]{inputenc}
```

La première de ces lignes demande à LaTeX d'utiliser les conventions allemandes pour les dates, coupures de mots, et liens.

La deuxième ligne demande à LaTeX d'utiliser les fontes Latin Modern. Ces polices sont en accord avec les polices Computer Modern de Donald Knuth pour les 128 premières positions, mais renferment des accents supplémentaires, trémas, points d'interrogation culbutés..., utilisés en Europe occidentale.

La troisième ligne permet à LaTeX d'établir la connexion entre les caractères du fichier et les glyphes eux-mêmes (c'est-à-dire, la représentation physique des caractères dans le document final).

Comme expliqué ci-dessus, la dernière ligne indique à LaTeX le codage qui a été utilisé dans le fichier source.

Les utilisateurs qui veulent plus de détails devraient consulter les documentations de babel, lmodern, et fontenc en utilisant l'élément Afficher l'aide pour le package... du menu Aide de TeXShop. La documentation est intéressante car elle retrace, dans une large mesure, l'historique de l'évolution de la conception d'une police dans TeX.

8 Perception du codage par TeXShop.

Le tableau ci-dessous montre, pour les codages les plus connus utilisés avec LaTeX dans TeXShop, les correspondances entre les types d'entrées.

La colonne « dialogues Ouvrir... / Enregistrer... » montre la dénomination du codage dans les dialogues de TeXShop Ouvrir... / Enregistrer... ; vous pourriez avoir besoin de cliquer sur le bouton Options pour afficher le menu déroulant des codages.

La colonne « directive de codage » montre la dénomination utilisée dans la directive de TeXShop,

```
% !TEX encoding = xxxxx
```

où xxxxx est l'indication du codage que vous souhaitez utiliser. Si cette ligne est déjà présente dans votre fichier source avant d'Enregistrer..., TeXShop sauvegardera automatiquement ce fichier dans le codage désigné. Sur un double-clic, TeXShop ouvrira également automatiquement le fichier avec

ce codage. Nous vous conseillons d'inclure cette directive dans vos modèles et de les utiliser pour créer vos nouveaux documents.

La colonne « inputenc » donne l'argument optionnel de l'extension inputenc. Comme avec la directive, je vous suggère de créer un modèle qui renferme la ligne inputenc avec le codage correspondant à celui indiqué dans la directive.

TABLEAU 1 – *Liste partielle des codages.*

TEXShop Dialogues Ouvrir… /Enregistrer…	TEXShop Directive de codage	LATEX inputenc
Unicode (UTF-8)	UTF-8 Unicode	utf8
Europe occidentale (Mac OS Roman)	MacOSRoman	applemac
Europe occidentale (ISO Latin 1)	IsoLatin	latin1
Europe centrale (ISO Latin 2)	IsoLatin2	latin2
Turquie (ISO Latin 5)	IsoLatin5	latin5
Europe occidentale (ISO Latin 9)	IsoLatin9	latin9
Mac Central European Roman	Mac Central European Roman	macee
Europe occidentale (Windows Latin 1)	Windows Latin 1	ansinew or cp1252