Gérer tout Unicode avec LaTeX ?

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 25 mars 2008

https://www.bortzmeyer.org/latex-unicode.html

Pour produire des documents de grande qualité typographique, surtout si on veut le faire de manière massive, sans passer par un cliquodrome nécessitant plusieurs clics par fichier, LaTeX reste imbattable. Mais si les documents contiennent de l'Unicode ?

LaTeX dispose depuis longtemps de moyens de formater du texte écrit dans différentes langues et différentes écritures. Par exemple, avec le paquetage **Babel**, on peut écrire des choses comme :

```
\usepackage[latin1]{inputenc}
\usepackage[french]{babel}
```

et typographier ainsi du français correct, directement saisi en Latin-1. Si on veut faire du grec, qui utilise un autre alphabet, ce n'est pas plus difficile :

```
\usepackage[greek]{babel}
...
\textgreek{...}
```

Mais ces techniques (décrites dans Langues exotiques [sic] <http://www.tuteurs.ens.fr/logiciels/latex/langues.html>) nécessitent de marquer le texte en connaissant sa langue. Elles ne reposent pas sur le modèle Unicode, où on peut représenter de manière non ambigüe tout caractère utilisé dans le monde. Écrire ces caractères en LaTeX n'est pas très difficile. On peut utiliser l'encodage UTF-8 ainsi :

```
\usepackage[utf8]{inputenc}
```

et taper ensuite de l'UTF-8. Mais cela ne règle que l'entrée dans le moteur TeX, pas la sortie, le rendu. En effet, les caractères non latins sont transformés en commandes LaTeX, mais pour un sous-ensemble d'Unicode seulement. Et ça ne marche pas à tous les coups. Ainsi, si le texte contient le caractère [Caractère Unicode non montré [1]] (`U+03A9`, grand oméga), LaTeX proteste :

```
...
(/usr/share/texmf-texlive/tex/latex/ucs/data/uni-33.def)
(/usr/share/texmf-texlive/tex/latex/ucs/data/uni-3.def)
! Undefined control sequence.
\u-default-937 #1->\textOmega

l.135 omega}" , ...
                  , (les symboles scientifiques ne sont normalement pas
?
```

Le grand oméga a été transformé en une commande LaTeX, `textOmega`... qui n'existe pas. (Ironie, le [Caractère Unicode non montré] - U+2126, Ohm - passe sans problème.) Pour afficher ce caractère grec, il faudrait installer des polices grecques comme `cbgreek` ou `cm-lgc` **et** marquer ces caractères avec une commande comme `textgreek`, plus haut.

Même chose avec des caractères cyrilliques (ici le [Caractère Unicode non montré], U+418, le grand I de Ivan) :

```
(/usr/share/texmf-texlive/tex/latex/ucs/data/uni-4.def)
! Undefined control sequence.
\u-default-1048 #1->\CYRI

l.98 programmes. En prenant l'exemple du ...

?
```

Pire, il n'y a pas de solution de repli, permettant de dire à LaTeX « si tu ne sais pas afficher un caractère, mets un joli carré blanc à la place ». Un caractère inexistant stoppe le programme LaTeX !

Si on tape directement en LaTeX, cela peut être supportable. Mais un programme qui recevrait de l'Unicode devrait décider d'une langue avant de pouvoir produire du LaTeX. Il devrait tester chaque caractère et faire quelque chose du genre (en pseudo-code) :

```
case groupe(le_caractère) is:

when greek => ...

when russian => ...

when arabic => ...
```

Des extensions complètement Unicode avaient été prévues comme Omega mais ont toutes été abandonnées. À l'heure actuelle, produire du LaTeX à partir de données contenant de l'Unicode quelconque est donc très pénible. C'est pour cela que la version PDF de mon blog <`https://www.bortzmeyer.org/pdf-version.html`> contient souvent un triste remplacement « Caractère Unicode non affiché ».

Merci à Erwan David, Ollivier Robert, Marc Baudoin, Bertrand Petit et Kim-Minh Kaplan pour leur aide et explications.

1. Car trop difficile à faire afficher par LaTeX