Finalement, je suis enfin passé à UTF-8

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 15 décembre 2012

https://www.bortzmeyer.org/passage-a-utf8.html

Normalement, en 2012, il y a longtemps que tout le monde est passé à UTF-8 et a abandonné le vieux Latin-1. Mais cela m'a pris plus longtemps que prévu et je viens juste de faire la transition.

Comme je l'avais expliqué dans un autre article <https://www.bortzmeyer.org/pas-encore-utf8.html>, c'est en effet plus compliqué que ça n'en a l'air, surtout si on a accumulé au cours du temps plein de fichiers, de programmes, et de réglages spécifiques à Latin-1. Et puis, ne nous voilons pas les yeux. S'il est normal aujourd'hui d'utiliser un encodage capable de représenter tous les caractères d'un coup, en pratique, il reste encore quelques trucs qui ne marchent pas bien en UTF-8. Pour donner une idée de l'opération, j'ai dû :

- Convertir les fichiers sources de ce blog (écrits en XML avec l'encodage Latin-1 <https://www.bortzmeyer.org/blog-implementation.html>) avec un petit script Python. (Pour un fichier individuel, c'est simple, on change l'encodage et le mode nxml-mode <http://www.thaiopensource.com/nxml-mode/> d'emacs convertit. Mais, ici, il y avait 1 500 fichiers...)
- Changer dans mon `/.zshrc` réglé à la main les variables environnement comme `LC_CTYPE`. Idem pour les fichiers de configuration d'autres logiciels comme le `.emacs` d'Emacs.
- Encore faut-il avoir un terminal qui affiche UTF-8. J'ai dû abandonner mon terminal favori, wterm, qui ne gère pas UTF-8 (il existe un wterm-ml mais qui a un exécutable différent par langue : pas pratique du tout). Par contre, xterm, gnome-terminal ou lxterminal <http://wiki.lxde.org/en/LXTerminal> se débrouillent très bien.
- Changer ses variables d'environnement ne convertit pas par magie les fichiers texte éparpillés un peu partout sur le disque. Pour certains répertoires importantes, j'ai tout converti à grands coups de commande recode <http://stackoverflow.com/questions/691040/converting-webpages-from-u
Pour les autres, je le ferai au fur et à mesure de leurs modifications.
- Changer la configuration de mutt (`set send_charset=us-ascii:utf-8`) et tester avec les répondeurs de courrier <https://www.bortzmeyer.org/repondeurs-courrier-test.html> pour être sûr que tout allait bien.
- Éditer `/etc/X11/fonts/misc/xfonts-base.alias` pour mettre une police Unicode pour X11.
- Remplacer des logiciels comme a2ps (qui ne gère pas du tout UTF-8 <http://bugs.debian.org/180236>). Pareil pour enscript. Il reste u2ps <http://u2ps.berlios.de/> et uniprint <http://www.sput.nl/unicode/print-utf.html>, mais ils ne font hélas pas de *"pretty-printing"*. (Il existe une liste des applications non-UTF8 sur Debian <http://wiki.debian.org/UTF8BrokenApps>.)

— Passer de l'ack-grep <http://betterthangrep.com/> un peu partout pour trouver tous les occurrences de « 8859 » ou de « Latin-1 » que j'aurais pu oublier. Cela ne m'a pas empêché d'en rater certaines et d'envoyer un fichier Latin-1 étiqueté UTF-8 sur la liste Frnog <http://www.frnog.org/>...

Voyons le côté positif, je peux désormais utiliser directement et nativement tous les caractères UTF-8 non Latin-1 comme [Caractère Unicode non montré [1]], œ, sans compter [Caractère Unicode non montré] ou des fantaisies comme [Caractère Unicode non montré]po[Caractère Unicode non montré][Caractère Unicode non montré]u[Caractère Unicode non montré] <https://www.bortzmeyer.org/unicode-envers.html>...

1. Car trop difficile à faire afficher par LaTeX