Unicode dans le Sudoc

Avant 2003, le Sudoc (Système universitaire de la documentation) était produit dans un environnement totalement « propriétaire » avec des données stockées en caractères Pica : par exemple, le « e accent aigu » était représenté par « \342e ». Si on voulait introduire un caractère accentué dans le système, il fallait donc, soit l'introduire directement sous cette forme-là, soit demander au système une conversion de la forme saisie vers cette forme-là. Il en allait de même dans l'autre sens : afficher ou exporter exigeait une conversion de la valeur Pica vers la valeur souhaitée (par exemple sa valeur ISO-8859-1 dans l'interface publique du catalogue).

Le nombre réduit de caractères Pica disponibles (latin-étendu incomplet) ne permettait pas d'envisager sérieusement une évolution dans ce contexte technique. C'est pourquoi la première opération, réalisée en 2003, a consisté à transporter le système central propriétaire [« CBS »] vers une plate-forme « ouverte » sous Unix.

La deuxième phase a consisté à permettre au système central de comprendre d'autres valeurs que celles des caractères Pica, les valeurs Unicode. Cette nouvelle version a été mise à disposition de l'Abes (Agence bibliographique de l'enseignement supérieur) en 2004. Parallèlement, il a fallu décider quelle forme d'Unicode (version 4) allait être utilisée par le système central : UTF-8, UTF-16, ou autre (8 bits, 16 bits, ou autre). UTF-8 (codage sur 1 à 4 octets, chacun sur 8 bits) a été choisi, parce que perçu comme plus « standard » à ce moment, et surtout parce que compatible avec les logiciels codant les caractères sur un octet. Il a également fallu statuer sur la forme d'UTF-8 qui serait le standard Sudoc. Il existe en effet deux façons de coder un « e accent aigu », par exemple : soit un « e » + un « accent aigu » (deux caractères, forme décomposée), soit un « e accent aigu » (un caractère, forme composée). Pica a choisi la forme décomposée. Ceci signifie que, en utilisant l'interface professionnelle et en tapant un « e accent aigu », on envoie dans le système \u0065 [= « e »] + \U00B4 [= « accent aigu »], et qu'il en sera de même en export UTF-8 standard. Ceci ne signifie pas que le système central ne comprenne pas \u00E9 [= « e accent aigu »], s'il le reçoit par copier/coller par exemple.

Rendre les clients compatibles

Une fois cette nouvelle version du système central installée, il ne restait plus qu'à rendre les « clients » compatibles, notamment pour les opérations de catalogage. Le premier a été le catalogue public (PSI, plus simple, parce que cette interface se contente de lire les données), opérationnel début 2005. Plus compliquée a été la mise en place de la nouvelle version de l'interface professionnelle (WinIBW, qui reçoit et envoie des données). Celle-ci a été mise à disposition au printemps 2005.

Par ailleurs, il a été décidé de ne pas convertir les données présentes dans la base [codées en caractères Pica], pour ne pas générer une augmentation importante et instantanée du volume total de la base (certains caractères Unicode en UTF-8 étant codés sur 4 octets). Depuis cette date donc, tant qu'une notice n'est pas modifiée, elle reste stockée dans la base en caractères Pica et convertie en Unicode si on l'appelle à l'affichage ou si on l'exporte. Elle ne devient une notice stockée en Unicode qu'à partir du moment où elle est modifiée dans la base. Ceci est complètement transparent pour les utilisateurs (professionnels ou publics), pour qui elle apparaît toujours comme une « notice Unicode ». Pour que la recherche ne soit pas affectée par la présence simultanée de caractères Pica et Unicode, tous les index, depuis la mise en service du nouveau système central, sont exclusivement en Unicode [ce qui permet de différencier une recherche sur « élève » et « élevé », par exemple].

Cataloguer dans toutes les écritures

Depuis l'été 2005, il est donc théoriquement possible de cataloguer un document en n'importe quelle écriture (connue de la version 4 d'Unicode). Cela suppose quand même la présence, sur le poste de travail ou de consultation, d'une police qui soit capable de restituer visuellement l'information [arial MS Unicode, par exemple], et d'outils permettant de produire les caractères des écritures autres que latines [les IME (claviers virtuels) de Microsoft, par exemple].

Après cela, il ne restait plus qu'à régler quelques menus détails comme l'affichage des écritures droite/gauche et les questions de translittération automatique, puisque la double saisie « écriture originale/écriture latine » est obligatoire dans le Sudoc.

2006 a été une année de travail sur ces deux questions : l'affichage des écritures droite/gauche n'est actuellement (relativement) satisfaisant que s'il n'y a pas de changement de sens d'écriture dans la même zone Unimarc. La translittération automatique est fonctionnelle, mais doit être améliorée pour coller aux normes de translittération française, et c'est le chantier 2007.

Christian Chabillon
Agence bibliographique de l'enseignement supérieur
chabillon@abes.fr