Unicode à ses débuts

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 20 octobre 2019

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

Sur le site Web du consortium Unicode, consortium qui pilote l'évolution de cette norme, on trouve une très intéressante page d'histoire <http://unicode.org/history/>, rassemblant un certain nombre de documents sur le passé d'Unicode. Parmi eux, « "*Unicode 88*" <http://unicode.org/history/unicode88.pdf> » qui, en 1988, était le premier document à exposer les bases de ce qui allait devenir la norme Unicode. Il est amusant de voir aujourd'hui ce qui a été gardé de ce projet, et ce qui a été changé.

Ce document « "*Unicode 88*" » était un projet : tout n'était pas encore défini. Mais on y trouve déjà quelques principes importants qui ont été conservés :

— Encodage de caractères abstraits, et pas de glyphes (la représentation graphique du caractère),
— Utilisation pour le texte brut, sans caractères de formatage (ce qui n'a pas été complètement respecté).

D'autres principes étaient absents, comme la séparation de l'affectation des points de code et leur encodage en bits. Il y a aussi des principes qui ont été gardés mais qu'on regrette aujourd'hui comme l'unification Han.

Le plus frappant, avec le recul, est l'insistance sur un principe qui n'a pas été conservé, l'encodage de taille fixe (proche du futur UCS-2), en 16 bits (dérivé de XCCS). On sait qu'en fait c'est un encodage de taille variable, UTF-8 (RFC 3629 [1]), qui s'est imposé (pour les protocoles Internet, ce choix en faveur d'UTF-8 est formalisé dans les RFC 2277 et RFC 5198.) L'article « "*Unicode 88*" » accuse son âge lorsqu'il écrit que 16 bits seront suffisants dans tous les cas (16 bits permettent d'encoder 65 536 caractères, or il en existe aujourd'hui 137 994 <https://www.bortzmeyer.org/unicode-12-0.html>). « "*Unicode 88*" » note que cela implique d'abandonner les écritures du passé, qui sont au contraire aujourd'hui une part importante d'Unicode.

À l'époque, un certain nombre de gens critiquaient Unicode en raison de l'augmentation de taille des textes (un caractère sur deux octets au lieu d'un). L'argument a toujours été faible, compte tenu

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc3629.txt

de la rapide augmentation des capacités des processeurs et des disques durs mais, à cette époque où la disquette de trois pouces et demi était une invention récente, il avait du poids. D'ailleurs, encore aujourd'hui, à une époque de documents Word de plusieurs dizaines de mégaoctets, sans même parler des vidéos haute définition, on entend parfois des critiques d'UTF-32 (un autre encodage de taille fixe…) sur la base de la taille des fichiers de texte brut. Le document estime que le problème de la taille est mieux réglé par la compression.

Autre point où le document accuse son âge, le curieux tableau qui mesure l'importance des écritures en fonction du PNB des pays qui les utilisent. Cela rappele qu'Unicode a été conçu par des entreprises capitalistes qui cherchaient le profit, et donc les marchés rentables.

Le choix d'un encodage de taille fixe (qui n'a donc pas été retenu par la suite) était stratégique, et le document revient plusieurs fois sur l'importance de ce choix, en raison de la simplification des programmes qu'il permet. L'argument de la simplicité des programmes a finalement cédé face à l'argument choc d'UTF-8 : la compatibilité avec ASCII (tout document en ASCII est automatiquement un document en UTF-8).

Et l'unification Han ? Cette idée de considérer les caractères chinois et japonais comme équivalents, en raison de leur origine commune, et malgré leurs apparences différentes (mais rappelez-vous qu'Unicode encode des caractères, pas des glyphes), est à peine discutée dans l'article, alors qu'elle est certainement le concept Unicode le plus critiqué depuis.