Les textes

Jean-Christophe Dubacq

S1 2016

1 Les textes

1.1 De l'écrit au binaire

1.1.1 La table ASCII

Trouvez dans la table ASCII :

1. Le caractère de code 0x41
2. Le caractère de code 0x30
3. Le caractère *a* et *A*. Comparez l'écriture binaire des codes numériques correspondants.
4. Le caractère de code 0x20. Quel est-il ?
5. Le caractère *retour chariot* (son nom est NEWLINE ou NL).

Comment passe-t-on d'une lettre à la suivante ? D'une majuscule à une minuscule ?

1.2 Jeux de caractères et codages

1.2.1 Codage nationaux et Mojibake

Soit le texte : Coefficient marée trop fort pour livraison tomates cœur-de-bœuf

1. Identifiez dans ce texte les ligatures linguistiques et les ligatures esthétiques
2. Est-il possible de représenter ce texte dans le jeu de caractères ASCII ?
3. Dans le jeu de caractère ISO-8859-15 (dit *latin-9*), il est possible de coder ce texte. Chaque caractère est alors codé par un octet unique. Quelle est la taille du fichier qui contient uniquement ce texte ?

4. Un polonais lit sur son vieil ordinateur le texte précédent. Il voit qu'une des lettres a été remplacée par ˝ (c'est un double accent aigu, comme dans Erdős, et pas un tréma comme dans Gwenaël). Laquelle et pourquoi ? S'il renvoie le texte tel quel a son correspondant français du début, que verra le français et pourquoi ?

1.2.2 UTF8

1. Le caractère de numéro 0x0041 (A) est codé par quel(s) octet(s) en UTF-8 ?
2. Le caractère de numéro 0x00E9 (é) est codé par quel(s) octet(s) en UTF-8 ?
3. Le caractère de numéro 0x0F03 () est codé par quel(s) octet(s) en UTF-8 ?
4. Le caractère de numéro 0x12084 () est codé par quel(s) octet(s) en UTF-8 ?
5. Dans un fichier codé en UTF-8, on trouve les six octets suivants. Combien de caractères sont réellements codés dans ce texte ?
6. L'anglais n'utilise que des caractères dont le numéro est dans la première ligne, et est codé traditionnellement en ISO-8859-1 (1 caractère = 1 octet). Le français utilise 5% de caractères de la deuxième ligne (le reste de la première), et est codé pareil (1 caractère = 1 octet). L'arabe (le russe, l'hébreu, le grec) sont aussi codés traditionnellement par 1 caractère = 1 octet, et comportent 95% de caractères de la deuxième ligne (le reste de la première ligne). Le chinois, en revanche est traditionnellement codé en BIG5 (1 caractère = 2 octets). Les textes chinois sont à 99% des caractères de la troisième ligne (le reste de la première ligne).

   Pour un texte de 1000 caractères codé en UTF-8, combien d'octets seront utilisés en moyenne pour un texte anglais, français, russe et chinois ?
7. Quel est en chinois l'augmentation de la taille du texte par rapport au codage traditionnel ?

1.3 Les chaînes de caractères

1.3.1 Les échappements en C

Dessinez quelle est la structure en mémoire des chaînes C suivantes ? Comment sont elles affichés ?

1. `"Toto"`
2. `"Bonjour le monde\n"`
3. `"Acheter:\n\tponey\n\tporte-avions\n"`
4. `"\303\251\n"`
5. `"\U20AC"` (symbole euro)
6. `"\0"`


Une bizarrerie historique du C/C++ fait que certaines séquences sont remplacées avant compilation par d'autres caractères :

Trigraphe	??(	??)	??<
Remplacement	[	]	{

7. `"Hello??!"`
8. `"Bye??/n"`

.1 La table ASCII