Representation de texte - Cours

– Février 2023

Dans cette séquence, on traitera des différentes façons de faire correspondre un caractère (une lettre, un espace...) avec une représentation binaire.

1 Table et encodage

Définition : Table ASCII

En 1960, la norme **ASCII** (American Standard Code for Information Interchange) a été crée pour uniformiser l'encodage des caractère par **l'ANSI** (American National Standards Institute).

Elle définit 128 codes, comprenant 95 caractères imprimables :

- les chiffres arabes de 0 à 9
- les 26 lettres de l'alphabet latin en minuscules et en capitales
- des symboles mathématiques et de ponctuation

Chaque caractère est codé sur 7 bits même si on le représente le plus souvent sur 8 bits (1 octet)

Remarques :

- Avantages de la table ASCII : .....
- Limitations : .....

Définition : ISO-8859-1 ou Latin-1

Pour rendre l'utilisation de la table ASCII plus universelle d'autres tables ont été créées.

L'ISO (Organisation internationale de normalisation) a proposé la norme ISO-8859 qui utilise le 8e bit pour ajouter 128 caractères supplémentaires pour un total de $2^8 = 256$ caractères.

Parmi les tables issues de cette norme, la table ISO-8859-1 (ou Latin-1) est celle qui a été le plus utilisée en occident car elle ajoute les caractères accentués et des nouveaux signes de ponctuation.

Remarques :

- Avantages de la table ISO-8859-1 : .....
- Limitations : .....

Définition : Unicode

Pour assurer l'universalité de la représentation de caractères la norme **Unicode** découpe l'encodage en deux étape :

- Le point de code : association entre un caractère et un **point de code** codés sur 20 ou 21 bits
- L'encodage du point de code (UTF-n où n est le nombre minimal de bit pour représenter un point de code).

Encodages les plus utilisés :

- UTF-8 : le point de code est encodé sur 1 à 4 octets (ou 8 à 32 bits)
- UTF-16 : le point de code est encodé sur 2 à 4 octets (16 à 32bits)
- UTF-32 : le point de code est encodé sur 4octets (ou 32 bits)

Encodage et décodage avec l'Unicode (en vous aidant de `https://unicode-table.com/fr`)

Caractère	Point de code	UTF-8	UTF-32
a			
€			
ℵ			

Remarques :

- Avantages de l'Unicode : .....
- Limitations : .....