Codage du texte

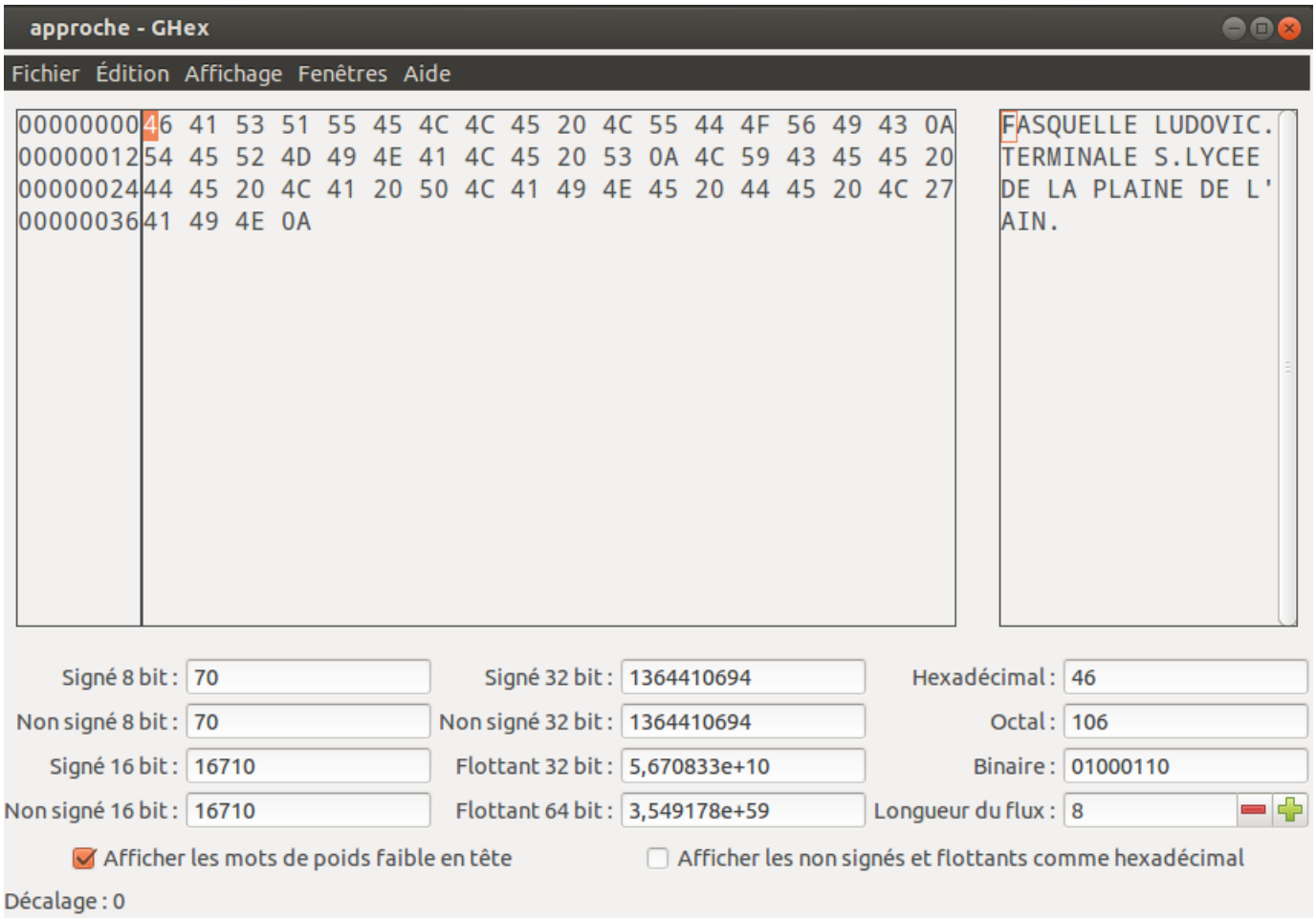
1 Approche du codage

Un ordinateur ne manipule que des 0 et des 1. Comment alors code-t-il du texte ?
Ouvrir un editeur de texte Kate par exemple :

- Ecrire en majuscule votre PRENOM, un espace puis votre NOM puis aller à la ligne.
- Ecrire TERMINALE S puis aller à la ligne
- Ecrire le nom du lycée toujours en Majuscule.

Enregistrer ce texte en tant que texte brut sans extension sous linux, txt sous windows).

Ouvrir ce fichier dans un éditeur hexadécimal GHex. Pour faciliter la lecture, on utilise ici l'hexadécimal



On peut lire dans la fenêtre de gauche les codes binaires correspondant aux caractères écrits. Dans la fenêtre de droite on lit le texte écrit dans l'éditeur de texte.

— Aux débuts (En 1963, la première version publiée de l'ASCII apparaît), on a établit une correspondance unique entre chaque caractère et un entier naturel : le **Charset**.
— L'ordinateur ne comprenant que le binaire, il a donc fallu représenter ces codes par des octets : l'**Encoding**.
— On a utilisé le code ASCII (American Standard Code for Information Interchange) qui est toujours en usage.

Sachant que l'on dispose de 127 caractères en ASCII, combien de bits sont nécessaires pour codés tous les caractères du code ?
Le 8ème bit sert de contrôleur d'erreurs, fréquentes avec les premières mémoires électroniques.

Vérifier à l'aide de la table ASCII : (Table ASCII) que le codage que l'on a obtenu correspond bien au texte tapé.

A savoir

Au commencement, chaque caractère était identifié par un code unique qui est un entier naturel et la correspondance entre le caractère et son code était appelée un Charset. Le code n'étant pas utilisable tel quel par un ordinateur qui ne comprend que le binaire, il fallut donc représenter les codes par des octets, et cela fut appelé Encoding.

Dans de nombreux grimoires anciens on découvre le code ASCII qui était utilisé pour représenter du texte en informatique. ASCII signifiait American Standard Code for Information Interchange. Il paraît que ce code est toujours en usage...

Le code ASCII se base sur un tableau contenant les caractères les plus utilisés en langue anglaise : les lettres de l'alphabet en majuscule (de A à Z) et en minuscule (de a à z), les dix chiffres arabes (de 0 à 9), des signes de ponctuation (point, virgule, point-virgule, deux points, points d'exclamation et d'interrogation, apostrophe ou quote, guillemet ou double quotes, parenthèses, crochets etc.), quelques symboles et certains caractères spéciaux invisibles (espace, retour-chariot, tabulation, retour-arrière, etc.).

Les créateurs de ce code limitèrent le nombre de ses caractères à 128, c'est-à-dire 2^7 , pour qu'ils puissent être codés avec seulement 7 bits : les ordinateurs utilisaient des cases mémoires de un octet, mais ils réservaient toujours le 8e bit pour le contrôle de parité (c'est une sécurité pour éviter les erreurs, qui étaient très fréquentes dans les premières mémoires électroniques).Les caractères sur 7 bits sont donc numérotés entre 0 et 127 (7F en hexadécimal, 0111 1111 en binaire).

Exemple : Le caractère A est codé en ASCII par le nombre 65 (dans notre système décimal habituel

2 La taille d'un texte

Quelle est la taille (en octets) de la phrase : ≪Enfin ! Je viens de comprendre ce qui s'est produit. ≫(attention, il faut compter les espaces, et signes de ponctuation…) ?

Vérifiez en tapant cette phrase avec un éditeur de texte quelconque. Il suffit d'écrire le texte, puis de l'enregistrer en tant que texte brut et ensuite de vérifier la taille en octets du fichier obtenu.

On peut ensuite écrire la même chose dans un logiciel de traitement de texte (comme LibreOffice Writer) et se rendre compte que la taille du fichier obtenu n'est pas du tout la même. Quelle peut en être l'explication ?

3 Utilisation de la table ASCII

1. A l'aide de la table ASCII, coder en binaire la phrase suivante : ≪L'an qui vient ! ≫.
2. Voici maintenant une exclamation codée en binaire :
   01000010 01110010 01100001 01110110 01101111 00101100
   Retrouver cette exclamation !
3. Peut-on coder en binaire la phrase ≪Un âne est-il passé par là ? ≫à l'aide de la table ASCII ? (Justifier la réponse)
4. Pour étendre la table ASCII, on utilise le 8ième bit pour coder les nouveaux caractères car on a développé de nouvelles méthodes de contrôle d'erreurs et les mémoires sont devenues plus fiables. Ce jeu de code est connu sous l'appellation ANSI (Amerian National Standards Institute) Combien de nouvelles possibilités de codage obtient-on alors ?

   On peut ainsi coder des lettres accentuées (présentes en français et en espagnol par exemple), des caractères typographiques utiles comme des tirets de diverses tailles et sortes, les caractères é, è, ç , à, ù, ô, oe, æ qui ne figurent pas dans la table ASCII.

4 la difficulté de la multiplication des normes

Le bit supplémentaire n'a pas permis encore de prendre en charge tous les caractères. Il a donc fallu mettre en place de nouvelles normes.

http ://fr.wikipedia.org/wiki/ISO$_8$859 − 1

http ://fr.wikipedia.org/wiki/Windows-1252

http ://fr.wikipedia.org/wiki/ISO$_8$859 − 15

Un même caractère pouvant être codé différemment suivant la norme utilisée, c'est une source de grande confusion pour les développeurs de programmes informatiques

Voici le code binaire d'un texte :

1. A l'aide de l'éditeur hexadécimal, régler en mode binaire, retrouve le texte contenu dans le code.
   01000101 01101110 00100000 01001001 01010011 01001110 00101100 00100000 01101001 01101100 00100000 01100110 01100001 01110101 01110100 00100000 01110010 01100101 01101110 01100100 01110010 01100101 00100000 01110011 01101111 01101110 00100000 01110100 01110010 01100001 01110110 01100001 01101001 01101100 00100000 01110011 01100001 01101110 01110011 00100000 01110010 01100101 01110100 01100001 01110010 01100100 00100001 00100001 00100001
2. Déterminer une astuce pour connaître la norme (ISO 8859-1,ISO 8859-15) utilisée par ce logiciel.

5 Quand le net s'affole

On a tous recu un jour un courriel bizarre ou lu une page web tellle que celle ci :

Prenons lâ€™exemple typique de la lumiÃ¨re Ã©mise par un phare maritime : elle est dâ€™abord indivisible, son coÃ»t de production Ã©tant alors indÃ©pendant du nombre d'utilisateurs ; elle possÃ¨de une propriÃ©tÃ© de non-rivalitÃ© (elle ne se dÃ©truit pas dans l'usage et peut donc Ãªtre adoptÃ©e par un nombre illimitÃ© d'utilisateurs) ; elle est Ã©galement non excluable car il est impossible dâ€™exclure de lâ€™usage un utilisateur, mÃªme si ce dernier ne contribue pas Ã son financement.

Quelle explication peut on donner à ce problème.

6 L' Unicode

La globalisation des échanges culturels et économiques conduit à la coexistence de nombreuses langues : les langues européennes et de nombreuses autres langues aux alphabets spécifiques voire sans alphabet. L'emploi d'internet dans le monde entier a donc nécessité la prise en compte d'un nombre plus important de caractères (par exemple, le mandarin possède plus de 5000 caractères).

D'autre part, avec le faible nombre de caractères pris en compte, cela pouvait conduire à des confusions : tous les symboles monétaires ne sont pas tous représentés dans le système ISO 8859-1 d'où des incompréhensions préjudiciables dans les ordres de paiement par courrier électronique.

Un consortium composé d'informaticiens, de chercheurs, de linguistes et de personnalités représentant les Etats ainsi que les entreprises s'occupe d'unifier toutes les pratiques en un seul et même système : la norme UNICODE.

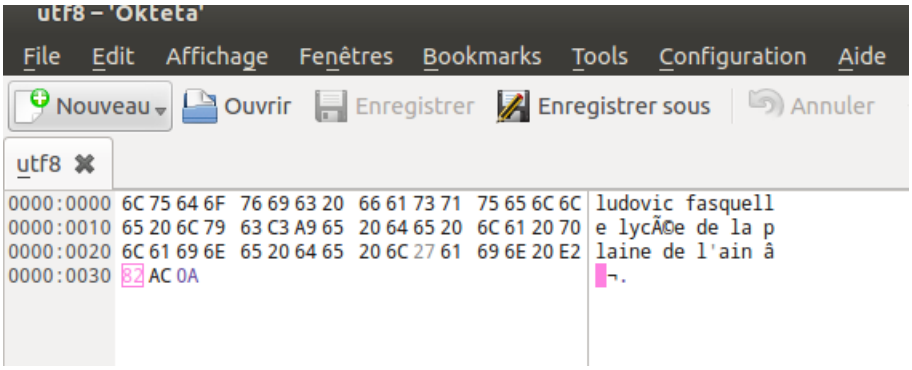
L'Unicode est la table de correspondance Caractères ↔ code (Charset).
L'UTF-8 est l'encodage correspondant (Encoding) le plus répandu (par défaut, les navigateurs Internet utilisent le codage UTF-8), il permet de coder tout caractère indépendamment de tout système de programmation ou système d'exploitation.

Ce code est utilisé dans les pages WEB, les emails, etc... Chaque caractère est représenté sous la forme d'un bloc U+xxxx (où xxxx est un nombre hexadécimal de 4 à 6 chiffres, entre U+0000 et U+10FFFF).La plage définie permet d'attribuer jusqu'à 1 114 112 points de code.

Par exemple les points de code U+0000 à U+FFFF contient la plupart des caractères utilisés par les langues modernes les plus courantes dans le monde.

Ecrire votre Prénom, Nom, le nom du lycée etc ... en miniscule en mettant les accents puis ajouter le symbole € et sauvegardez avec un encodage en utf-8.

Ouvrir le document précédent avec l'éditeur héxadecimal et examiner les codes.



Repérez les 3 octets appelé BOM (Byte Order Mark) indiquant le codage UTF-8.

Repérez les 2 octets utilisés pour le caractère é

Repérez les trois octets utilisés pour le caractère €

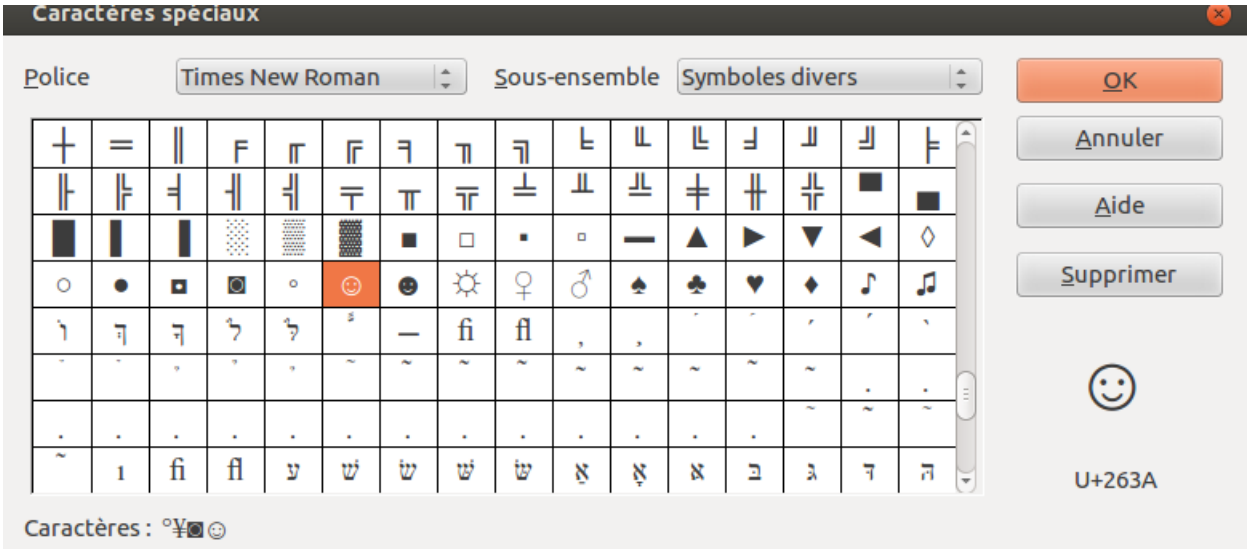
En utilisant la fiche ressource sur la norme UTF-8, expliquer les codages précédents, détailler le codagede €.

7 Monsieur Jourdain et l'UNICODE

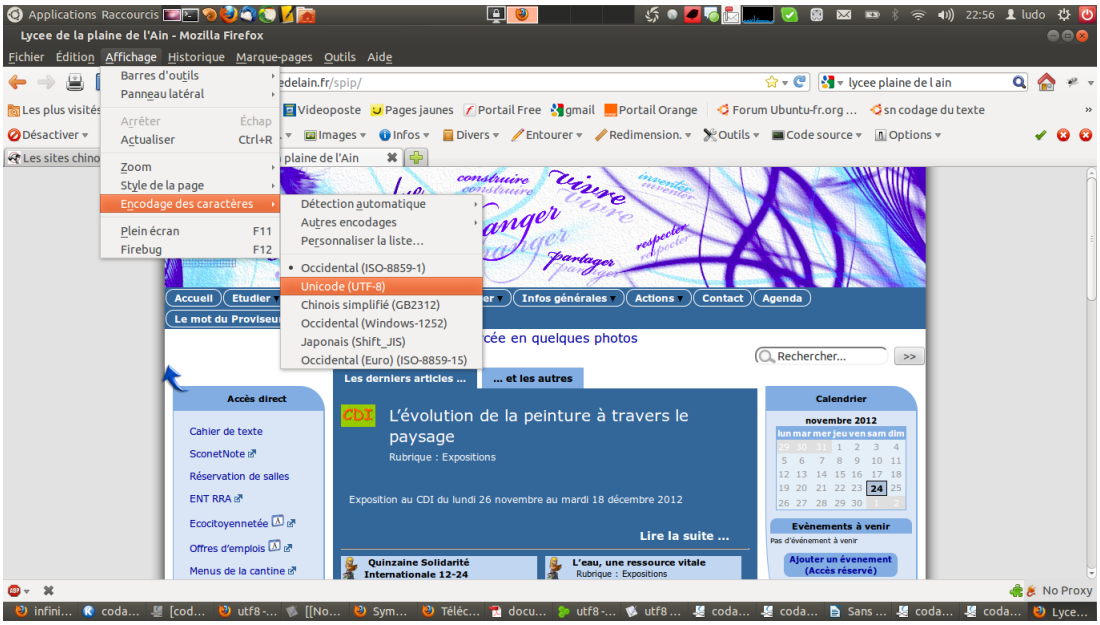
Monsieur Jourdain faisait de la prose sans le savoir et vous ?? vous utilisez l'Unicode sans le savoir.

7.1 Traitement de texte libreoffice

Menu Insertion > caractères spéciaux et cliquer sur un caractère qui ne peut pas s'écrire au clavier et vous aurez le codage Unicode associé.



7.2 Dans les pages web (Firefox) ou les clients de messagerie



Aller sur un site chinois et observez.Vous pourrez utiliser plus d'encodage pour résoudre les éventuelles problèmes. Pour des compléments : http ://fr.wikipedia.org/wiki/UTF-8

8 Programme Python

— Ecrire un programme python qui convertit le code binaire en texte.
En entrée on aura : textebin=['01000010', '01110010', '01100001', '01110110', '01101111', '00101100'] et en sorie le texte correspondant.

— Ecrire un programme python qui convertit le texte en code binaire.
En entrée on aura : texte=texte="L'an qui vient !" et en sorie le code binaire correspondant.