- Représentation du texte

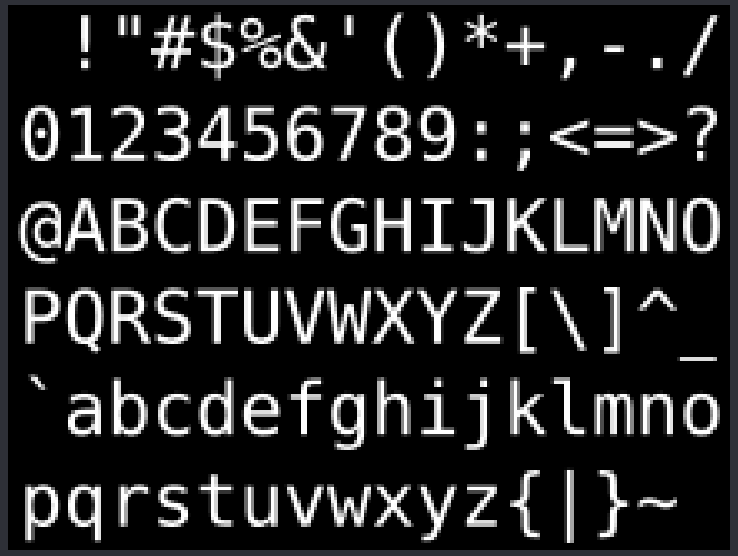
Contenu
Représentation du texte

a. Le code ASCII

b. Taille et format du fichier texte

Un fichier texte

- Un **fichier texte** ou **fichier texte brut** ou **fichier texte simple** est un fichier dont le contenu représente uniquement une suite de caractères; il utilise nécessairement une forme particulière de codage de caractère qui peut être une variante ou une extension du standard local des États-Unis, l'ASCII.



Historique…1

-1950 Le mot fichier (« *file* » en anglais) a été utilisé publiquement dans le contexte d'un enregistrement informatique .

-Une publicité de la radio Radio Corporation of America de Popular Science Magazine décrivant une nouvelle mémoire à tube à vide qu'elle avait développée, expliquait :

« ...the results of countless computations can be kept "on file" and taken out again. Such a "file" now exists in a "memory" tube developed at RCA Laboratories. Electronically it retains figures fed into calculating machines, holds them in storage while it memorizes new ones - speeds intelligent solutions through mazes of mathematics. »

-1952 Un fichier était utilisé pour désigner l'information enregistrée sur une carte perforée.

Historique …2

-1971 le RFC 265 indique qu'un fichier peut être ASCII, cœur d'exécutable, ou autre. Il mentionne notamment l'EBCDIC.

-1972 le RFC 354,discutant des échanges de texte par le protocole réseau NVT-ASCII FTP, indique que les fichiers textes sont enregistrés de manières différentes selon les systèmes.

Historique …3

-Le PDP-10 enregistre le NVT-ASCII en 7-bits justifiés à gauche dans des mots de 36 bits.

-Le 360's enregistre le texte avec un codage EBCDIC 8-bit.

-Multics enregistre le texte avec quatre caractère de neuf bits dans des mots de 36 bits.

Historique…4

Il indique donc que pour le bon transfert des textes, il est nécessaire que les deux parties effectuent leur part respective de la conversion dans un codage commun; il s'agit à cette époque de l'ASCII 8 bits, dit NVT.

-Cette même année, la RFC fait apparaître le besoin d'une unité commune pour transmettre des données binaires entre systèmes dont les mots n'ont pas la même taille, et suggère l'utilisation de byte de 8 bits, c'est-à-dire, de ce que l'on appelle aujourd'hui des octets.

Historique…5

-En 1980, le RFC 765 en spécifiant le protocole FTP indique les trois raisons occasionnant le transfert d'un fichier : l'impression, l'archivage, et le traitement.

-En 1985, lors de l'élaboration du protocole FTP de transfert de fichier, il a été recommandé de considérer comme fichier texte (en anglais "text" files), deux formats de fichiers :

-les fichiers file structure, où le fichier est considéré être une séquence continue de lignes.
-les fichiers record-structure, où le fichier est constitué d'enregistrements séquentiels.

C’est quoi l’apport??

-Le fichier texte, lorsqu'il apparait apporte la possibilité de permettre à un humain de soumettre un texte au traitement automatique d'une machine.

- Il offre également la possibilité de supprimer et d'ajouter une ligne, et cela dès les cartes perforées. Cette fonctionnalité a été reprise par des logiciels comme ed ou edlin .

Y a il des Limitations??

-Un fichier texte est limité dans sa taille, comme le sont tous les fichiers, par le système de gestion de fichiers.

-Le fichier texte peut poser de nombreux problèmes d'interopérabilité (pour cause d'encodage différents) entre pays, entre fournisseurs de logiciels, notamment.

Usage

-Les fichiers texte sont utilisés par de nombreux logiciels.

-Ils sont également utilisés pour contenir les textes écrits en langages de programmation. En outre, la plupart des langages de programmation offrent des fonctions prédéfinies pour manipuler du texte brut, ce qui rend la gestion des fichiers textes particulièrement accessible.

-Le logiciel utilisé pour éditer un fichier texte est un éditeur de texte.

-Dans le cas général, un traitement de texte ne produit pas des fichiers texte. En effet, un traitement de texte n'a pas seulement besoin de manipuler du texte brut, mais également des informations sur la fonte de caractère utilisée, la disposition des caractères dans des pages, les styles typographiques, etc.

Structure…1

La structure d'un fichier texte est une séquence de lignes. Toutefois, historiquement, chaque caractère est aligné verticalement, c'est encore le cas aujourd'hui, dans un éditeur de texte en ligne ou local.

-Séquence de lignes

-Le concept de séquence de lignes reste une caractéristique forte d'un fichier texte.
-Un fichier texte peut simplement contenir du texte dans une langue quelconque.

“Dance like there's nobody watching,
Love like you'll never be hurt.
Sing like there's nobody listening,
And live like it's heaven on earth.”

Structure…2

-Un fichier texte peut également contenir une donnée structurée qui peut être analysée par un logiciel et affichée sous une forme plus évoluée, par exemple une page web:

<!DOCTYPE html> <html lang="fr"> <head><title>Page web d'exemple</title></head> <body> Ceci est une page web d'exemple. </body> </html>

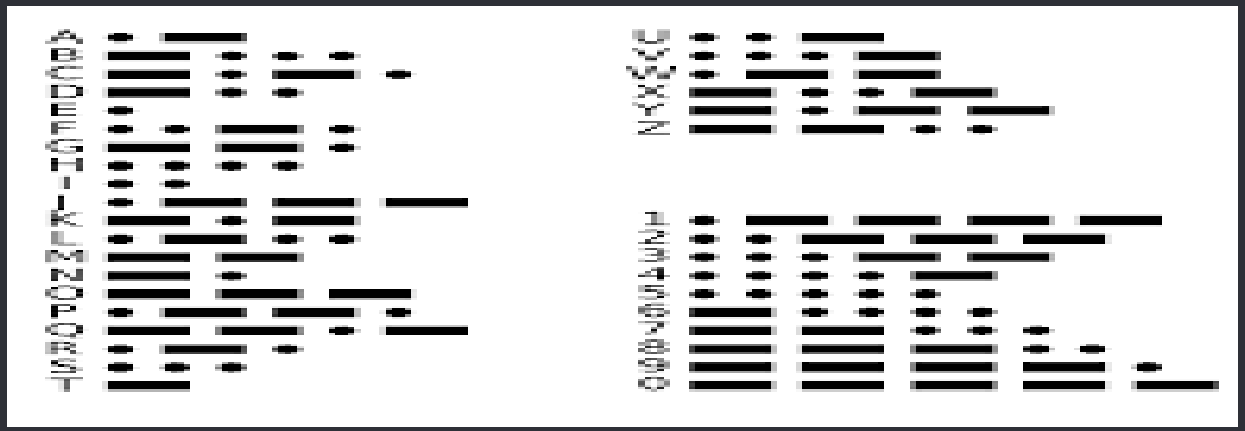
- *Code ASCII*

Historique…1

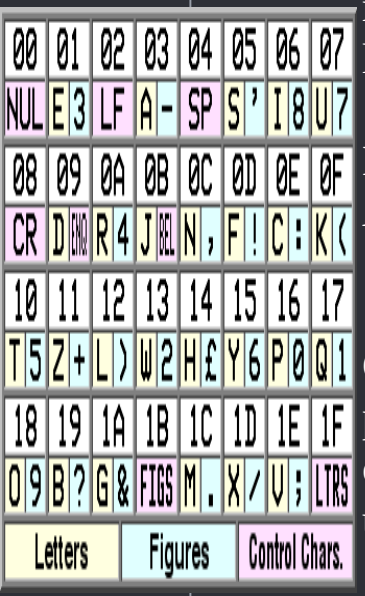
-1844 Le morse a été le premier codage à permettre une communication longue distance.

-Ce code est composé de points et de tirets (un codage binaire en quelque sorte...). Il permit d'effectuer des communications beaucoup plus rapides que ne le permettait le système de courrier de l'époque aux Etats-Unis : le Pony Express.

-De nombreux codes furent inventés dont le code d'Émile Baudot (portant d'ailleurs le nom de*code Baudot*, les anglais l'appelaient en revanche *Murray Code*).



Historique…2



-1876 le Dr Graham Bell met au point le téléphone, une invention révolutionnaire qui permet de faire circuler de l'information vocale dans des lignes métalliques.

La Chambre des représentants a décidé que l'invention du téléphone revenait à Antonio Meucci. Ce dernier avait en effet déposé une demande de brevet en 1871, mais n'avait pas pu financer celle-ci au-delà de 1874.

Ces lignes métalliques permirent l'essor des téléscripteurs, des machines permettant le codage et le décodage des caractères grâce au code Baudot (les caractères étaient alors définis sur 5 bits, il y avait donc 32 caractères uniquement...).

Dans les années 60, le code ASCII (American Standard Code for Information Interchange) est adopté comme standard. Il permet de coder les caractères sur 8 bits, soit 256 caractères possibles.

Qu'est-ce que le code ASCII ?

-La mémoire de l'ordinateur conserve toutes les données sous forme numérique.

-Il n'existe pas de méthode pour stocker directement les caractères. Chaque caractère possède donc son équivalent en code numérique : c'est le **code ASCII** (*American Standard Code for Information Interchange* - traduisez « Code Americain Standard pour l'Echange d'Informations »).

-Le code ASCII de base représentait les caractères sur 7 bits (c'est-à-dire 128 caractères possibles, de 0 à 127).

-Les codes 0 à 31 ne sont pas des caractères. On les appelle *caractères de contrôle* car ils permettent de faire des actions telles que :

-Retour à la ligne (CR)

-Bip sonore (BEL)

-Les codes 65 à 90 représentent les majuscules

-Les codes 97 à 122 représentent les minuscules

Table ASCII

caractère<b>	<b>code ASCII	code hexadécimal
NUL (*Null*)	0	00
SOH (*Start of heading*)	1	01
STX (*Start of text*)	2	02
ETX (*End of text*)	3	03
EOT (*End of transmission*)	4	04
ENQ (*Enquiry*)	5	05
ACK (*Acknowledge*)	6	06
BEL (*Bell*)	7	07
BS (*Backspace*)	8	08
TAB (Tabulation horizontale)	9	09
LF (*Line Feed*, saut de ligne)	10	0A
VT (*Vertical tabulation*, tabulation verticale)	11	0B
FF (*Form feed*)	12	0C
CR (*Carriage return*, retour à la ligne)	13	0D
SO (*Shift out*)	14	0E
SI (*Shift in*)	15	0F
DLE (*Data link escape*)	16	10
DC1 (*Device control 1*)	17	11

y	121	79
z	122	7A
{	123	7B
	124	7C
}	125	7D
~	126	7E
Touche de suppression	127	7F

Mais c’est quoi le ASCII Etendue??...1

-Le code ASCII a été mis au point pour la langue anglaise, il ne contient donc pas de caractères accentués, ni de caractères spécifiques à une langue.

-Pour coder ce type de caractère il faut recourir à un autre code. Le code ASCII a donc été étendu à 8 bits (un octet) pour pouvoir coder plus de caractères (on parle d'ailleurs de code ASCII étendu...).

-Ce code attribue les valeurs 0 à 255 donc codées sur 8 bits aux lettres majuscules et minuscules, aux chiffres, aux marques de ponctuation et aux autres symboles (caractères accentués)

-Le code ASCII étendu n'est pas unique et dépend fortement de la plateforme utilisée.

-Les deux jeux de caractères ASCII étendus les plus couramment utilisés sont :

-Le code ASCII étendu OEM, c'est-à-dire celui qui équipait les premières machines de type IBM PC

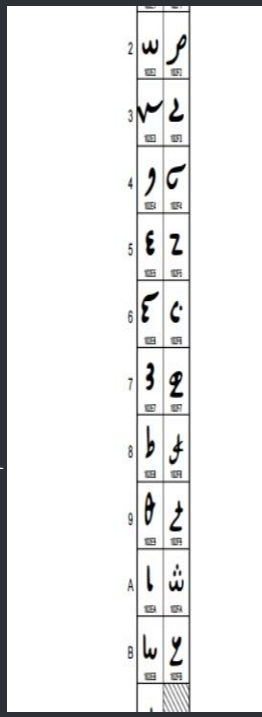
Mais c’est quoi le ASCII Etendue??...2

-Le code EBCDIC

-Le code *EBCDIC* (*Extended Binary-Coded Decimal Interchange Code*), développé par IBM, permet de coder des caractères sur 8 bits. Bien que largement répandu sur les machines IBM, il n'a pas eu le succès qu'a connu le code ASCII.

-Unicode

◦Le code *Unicode* est un système de codage des caractères sur 16 bits mis au point en 1991. Le système Unicode permet de représenter n'importe quel caractère par un code sur 16 bits, indépendamment de tout système d'exploitation ou langage de programmation.(La dernière version, Unicode 9.0, est publiée le 21 juin 2016)



-Il regroupe ainsi la quasi-totalité des alphabets existants (arabe, arménien, cyrillique, grec, hébreu, latin, ...) et est compatible avec le code ASCII.

Qu'est-ce qu'un format??

-Un format de données est une méthode d'écriture et de stockage des données utilisée en informatique pour représenter des données sous forme de nombres binaires.

-C'est une convention (éventuellement normalisée) utilisée pour représenter des données, autrement dit des informations représentant un texte, une page, une image, un son, un fichier exécutable, etc.

-Lorsque ces données sont stockées dans un fichier, on parle de format de fichier. Une telle convention permet d'échanger des données entre divers programmes informatiques ou logiciels,

-Un format peut être propriétaire, c'est-à-dire un format exclusif utilisé par un logiciel, ou non propriétaire c'est-à-dire ouvert. Nous traitons ici des deux (dans le tableau récapitulatif, vous pouvez voir quel format est propriétaire et quel format ne l'est pas).

Les Formats texte…1

			
Format PDF (Portable Document Format	Le format PDF, est un format de diffusion et de conservation normalisé par l'ISO sous les normes PDF/A-1 et PDF/X (respectivement ISO 19005-1 et ISO 15930). De ce fait, il est un format idéal pour transmettre des documents : sa mise en forme conservée quel que soit le logiciel ou le système d'exploitation de l'usager.	Le format PDF n'est pas modifiable aisément, de par son statut même : c'est un format de diffusion et non d'édition.	Le format PDF peut être ouvert : sous Windows, avec Adobe Reader sur MacOS, avec Aperçu sur Linux, avec xpdf
Format ODF (Open Document Format) et ses déclinaisons : odt (Text), ods (Spreadsheet), odp (Presentation)	Les formats OpenDocument sont une norme internationale pour la bureautique (ISO 26300:2006).	Le format OpenDocument n'est pas ouvrable avec des logiciels anciens.	La suite bureautique **OpenOffice** , depuis sa version 2.0, permet de créer, d'ouvrir et d'éditer des documents ODF dans le respect de la norme ISO. Elle est disponible sur Windows, Mac et Linux. Elle est téléchargeable gratuitement sur le site http://fr.openoffice.org. La suite Microsoft Office permet également, depuis sa version 2007 SP2, d'ouvrir et d'enregistrer les fichiers OpenDocument.

Les Formats texte…2

Formats (Office) Open XML (docx, xslx, pptx)	Les formats Open XML sont une norme internationale. Cependant, aucun logiciel ne respecte actuellement l'intégralité de la norme.	Les formats docx, xlsx et pptx ne sont pas ouvrables avec des logiciels anciens. Il existe néanmoins un plugin pour Office 2003, disponible sur le site de Microsoft.	La suite bureautique **Office 2007** de Microsoft (Word, Excel, PowerPoint) est celle qui se rapproche le plus de l'implémentation de la norme ISO. Depuis sa version 3, la suite gratuite OpenOffice permet également d'ouvrir et de modifier des documents créés avec le format Office Open XML. La suite OpenOffice peut être téléchargée gratuitement sur le site http://fr.openoffice.org.

Les Formats texte…3

Formats doc, xls, ptt	Ces formats sont lisibles par la majorité des traitements de textes / tableurs / logiciels de présentations du marché.	Formats obsolètes Aucune pérennité (jamais normalisé) Mise en forme boulversée à chaque logiciel / version Lourdeur du fichier (images mal compressées) Ne plus utiliser, sauf pour des soucis de compatibilité	Les formats doc, xls ou ptt s'ouvrent respectivement avec les outils Microsoft Word, Excel et PowerPoint. Les différentes versions de la suite OpenOffice savent également ouvrir ces formats.
Format rtf	Le format Rich Text Format est lisible par tous les traitements de texte . Standardisé par Microsoft	En déclin - plus maintenu Fichiers très lourds	

Les Formats texte…4

Texte brut (txt)	Lisible par tout éditeur de texte	Pas de mise en forme Pas de mise en page Pas prévu pour l'impression !	Le texte brut peut s'ouvrir avec n'importe quel éditeur de texte. Il est souvent associé au "bloc-notes" de Windows
Works (wps)	Format obsolète et abandonné Spécifique à l'outil Microsoft Works		Le format Works s'ouvrait avec la suite Microsoft Works. Il peut également être ouvert avec OpenOffice Writer.
OpenOffice 1.0 (sxw)	Format obsolète et abandonné (remplacé par OpenDocument) Spécifique à l'outil OpenOffice 1		Le format sxw peut être ouvert avec les suites OpenOffice et StarOffice.

Merci