Représentation d'un texte en machine

Afin de représenter tout type de caractères (lettre, symbole ou chiffre), il est indispensable d'utiliser un système de codage informatique, appelé encodage.
L'encodage assure la correspondance entre les caractères et les nombres binaires stockés dans la mémoire de l'ordinateur.
Un fichier informatique n'est finalement qu'une suite d'octets.
Utiliser tel encodage plutôt que tel autre revient à interpréter la même séquence d'octets comme 2 textes différents.

De nombreux codages coexistent, comme vue dans "Histoire des sciences".

- Le codage ASCII
  Simple mais limité en termes de caractères à encoder, ce qui ne le rend pas universel.
  7 bits suffisent pour coder un caractère.

- La norme ISO 8859-1 (ou Latin-1)
  Permet d'encoder tous les caractères des principales langues européennes.
  Chaque caractère est codé sur 1 octet.
  Certains alphabets comme le cyrillique ou le polonais ont leur propre norme.

- Le standard Unicode
  Il est universel et extensible si besoin.
  En revanche, il nécessite l'utilisation de 4 à 6 octets par caractère à encoder, ce qui allong la taille du message.

- La norme UTF-8
  C'est une représentation d'Unicode dont elle possède les avantages.
  Cet encodage est de taille variable, ce qui lui permet d'être moins gourmand en espace mémoire qu'Unicode.
  De plus, il est compatible avec ASCII, mais il est plus compliqué à gérer en machine.

Exemple :

Caractère	É	l	è	v	e
Codage UTF-8	C3 89	6C	C3 A8	76	65
ISO 8859-1	C9	6C	E8	76	65

Le nombre de symboles hexadécimaux étant variable, le nombre de caractères contenus dans un fichier texte n'est pas déductible à partir de la taille du fichier.
En Octobre 2020, 95% des sites web étaient encodés en UTF-8, ce qui permettait une uniformisation des pages et des navigateurs web.