Le codage de l'information numérique

Table des matières

I - Contexte 3

II - Principes du codage 4

III - Exercice : Appliquer la notion 6

IV - Discrétisation et numérisation de l'information 7

V - Exercice : Appliquer la notion 10

VI - Base numérique 11

VII - Exercice : Appliquer la notion 14

VIII - Représentation binaire 15

IX - Exercice : Appliquer la notion 18

X - Représentation des images bitmaps 19

XI - Exercice : Appliquer la notion 21

XII - Représentation du texte 22

XIII - Exercice : Appliquer la notion 24

XIV - Format de fichiers 25

XV - Exercice : Appliquer la notion 29

XVI - Essentiel 30

XVII - Quizz 32

Solutions des exercices 37

Crédits des ressources 47

I Contexte

Durée : 2h

Environnement de travail : Repl.it, terminal

Pré-requis : Aucun

On peut utiliser un ordinateur pour afficher des photographies, modifier un texte ou effectuer des calculs sur des nombres. Il paraît évident à un être humain qu'une image est un ensemble de formes ou de points, ou encore qu'un texte est une suite de lettres. Mais ce n'est pas une réalité accessible directement par l'ordinateur.

Il faut donc un moyen de représenter les informations quelles qu'elles soient (un pixel, une lettre, un nombre, etc.) d'une façon qui soit manipulable par l'ordinateur. On appelle cela le **codage** de l'information.

Comme un ordinateur ne sait manipuler que des nombres **binaires** (c'est à dire des séquences de 0 et de 1), il est nécessaire de représenter les informations que l'on souhaite manipuler par de telles séquences de 0 et de 1.

Ce module présente les principes du codage informatique :

- la **conversion** des informations analogiques en informations numériques (c'est à dire en nombres),
- la **représentation binaire** (la seule que la machine sait manipuler),
- les **formats** (qui permettent d'associer du sens aux séquences binaires).

Nous illustrerons ces concepts avec deux cas pratiques que nous rencontrons tous les jours sur nos écrans : le codage du texte et le codage des images.

II Principes du codage

Objectifs

- Découvrir la notion de codage ;
- Découvrir l'intérêt du codage ;
- Découvrir des exemples de codeur.

Mise en situation

Un ordinateur ne sait interpréter que des séquences de symboles. Afin d'interagir avec lui il faut donc trouver une manière de représenter les informations que l'on manipule en une forme interprétable et compréhensible par l'ordinateur. C'est l'objet du codage.

Codage Définition

« Le codage de l'information concerne les moyens de formaliser l'information afin de pouvoir la manipuler, la stocker ou la transmettre. Il ne s'intéresse pas au contenu mais seulement à la forme et à la taille des informations à coder.

http://fr.wikipedia.org/wiki/Codage_de_l%27information »

Le codage repose sur l'utilisation de **symboles** pour formaliser l'information.

Notes de musique et partition Exemple

Un accord en musique est composé de notes de musiques qui peuvent être vues comme des concepts posés sur papier via des symboles et notations, indiquant ainsi des informations comme la hauteur ou encore la durée de ces notes.

Le nombre douze : du concept aux symboles Exemple

Le nombre douze est un concept mathématique : il s'agit d'une abstraction pour désigner une certaine quantité d'objets.

Afin de réaliser des calculs à la main, par exemple, on le représente sous forme de symboles appelés chiffres '1' et '2' et on le note '12'.

Remarque

Il existe plusieurs codages pour le même concept, par exemple le nombre « douze » se code « XII » en chiffres romains.

À retenir

- Pour manipuler de l'information, on utilise un codage qui formalise cette information sous forme de symboles.

Exercice : Appliquer la notion

[solution n°1 p. 37]

Selon l'apocryphe de Camus :

« Mal nommer les choses c'est ajouter au malheur du monde. »

Dans la langue française, on utilise des concepts que l'on représente par des mots pour s'exprimer. On dispose de plusieurs codages avec chacun leurs symboles. Mettez en correspondance ces symboles et ces codages.

A Des phonèmes **B** Des lettres **C** Des gestes

On peut écrire les mots : on dispose donc d'un codage écrit composé de symboles qui sont	On peut dire les mots : on dispose d'un codage vocal composé de symboles qui sont	On peut mimer les mots : on dispose d'un codage — la langue des signes — composé de symboles qui sont

IV Discrétisation et numérisation de l'information

Objectifs

- Découvrir la notion de signal discret ;
- Découvrir la notion de convertisseur analogique numérique.

Mise en situation

Notre interaction avec le monde physique se fait via de nombreux signaux **continus** (comme le son ou la lumière), dits **analogiques**. Pour traiter ces signaux avec un ordinateur, il faut effectuer une conversion afin d'obtenir une représentation **numérique** : on parle de **discrétisation** et de **numérisation** de l'information.

La discrétisation consiste à découper le signal en petits morceaux (par exemple des pixels pour une image) et la numérisation consiste à associer à chacun de ces morceaux un nombre qui représente l'information (par exemple une couleur pour une image).

Échantillonnage et signal échantillonné

Az Définition

Un **signal échantillonné** est une représentation composée d'échantillons d'un signal analogique. On ne sélectionne qu'une partie du signal en enregistrant uniquement certaines de ses valeurs.

Le processus associé est appelé **échantillonnage**.

Synonyme : on parle aussi de **discrétisation** à la place d'échantillonnage.

Un film à bandes

Exemple

Si on prend un film enregistré par un cinématographe (une ancienne caméra), celui-ci peut-être vu comme un signal échantillonné de la réalité prise par la caméra. Les échantillons ici sont des images.

Numérisation et signal numérisé

Az Définition

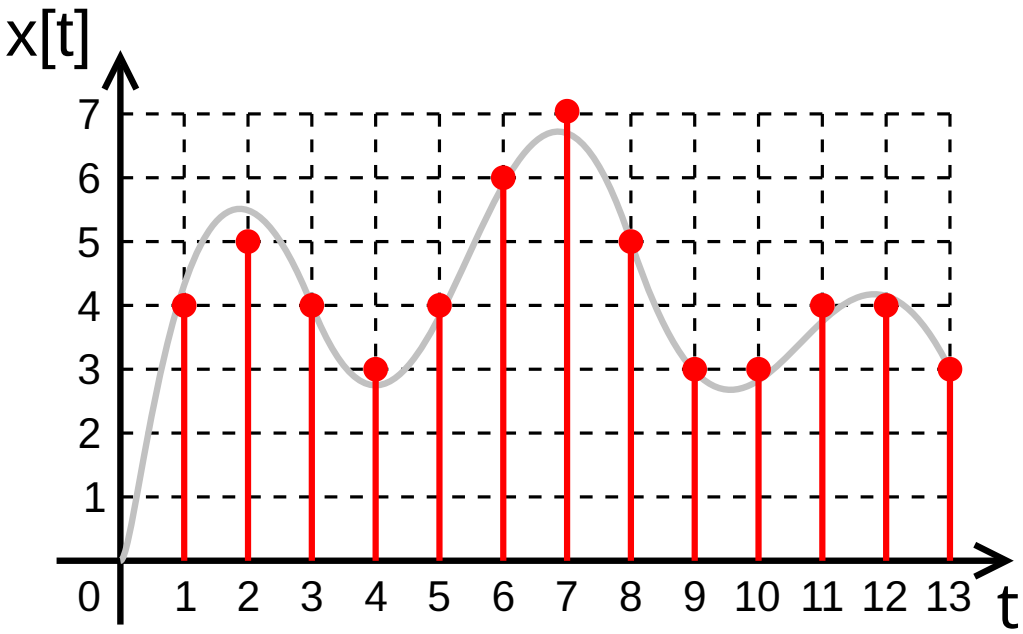
Après avoir obtenu des échantillons, on peut leur attribuer une valeur numérique : il s'agit du processus de **numérisation**.

Échantillonnage et numérisation du son

Exemple

Le son est un signal analogique : c'est une onde mécanique.

Pour l'échantillonner et avoir une version discrète de celui-ci, on sélectionne certaines valeurs à un intervalle de temps donné : une **période** temporelle.



Échantillonnage : Passage d'un signal analogique à un signal discrétisé

Après avoir réalisé cette discrétisation, on peut numériser le signal : pour chaque échantillon, on lui associe une valeur numérique.

À la fin de ce processus, on obtient un signal numérisé.

Convertisseur analogique numérique

Az Définition

Les systèmes qui réalisent l'échantillonnage et la numérisation des signaux analogiques sont appelés **convertisseurs analogique numérique**.

Microphone numérique : traduire un son en une version numérique.

Exemple

Un microphone numérique (par exemple sur un téléphone) convertit le signal physique, une onde, en une représentation numérique que l'on peut stocker et manipuler.

Codeur rotatif : traduire un angle en valeur numérique

Exemple

« Les codeurs rotatifs sont un type de capteurs permettant de délivrer une information d'angle, en mesurant la rotation effectuée autour d'un axe.

L'information de vitesse peut alors être déduite de la variation de la position par rapport au temps.

https://fr.wikipedia.org/wiki/Codeur_rotatif »

Codeur rotatif ROD 420

Ici une grandeur physique (analogique), l'amplitude d'un mouvement angulaire, est convertie en un nombre.

Les codeurs rotatifs sont par exemple utilisés par les ordinateurs de bord des véhicules : si on dispose du rayon des roues, on peut déduire la vitesse du véhicule.

Convertisseur analogique numérique ⊕ Complément

Pour en apprendre plus sur cette grande catégorie de système de codage, on pourra consulter l'article Wikipédia associé :

Convertisseur analogique numérique[1]

À retenir

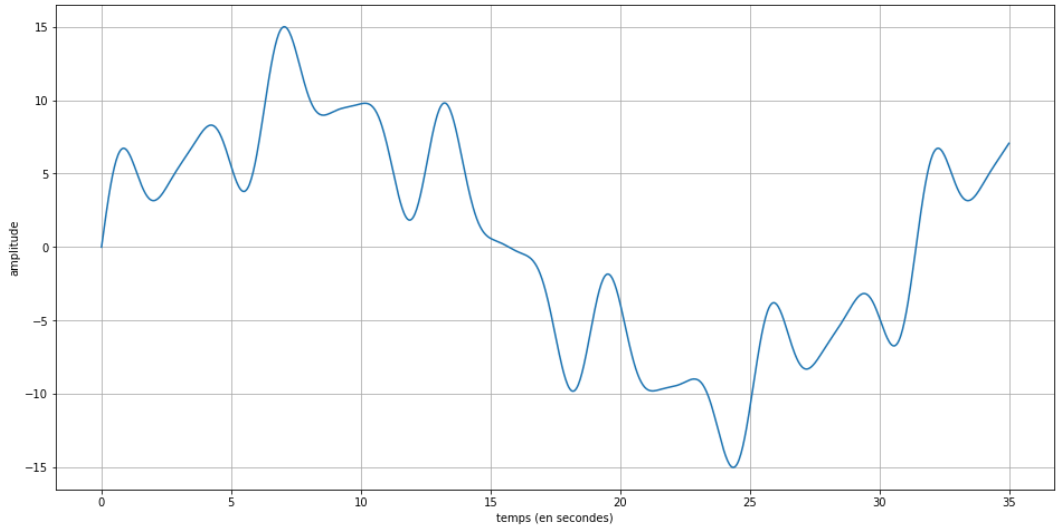
Les signaux analogiques – ceux du monde réel – sont convertis en signaux numérisés par l'intermédiaire de convertisseurs analogiques numériques qui réalisent :

- leur échantillonnage,
- et leur numérisation.

1. https://fr.wikipedia.org/wiki/Convertisseur_analogique-num%C3%A9rique

V Exercice : Appliquer la notion

On se donne la plage suivante d'un signal analogique observé.



Signal analogique

Question 1

[solution n°2 p. 37]

On souhaite le discrétiser. Pour cela, on choisit de prendre des échantillons de ses valeurs toutes les 5 secondes en prenant la première valeur à l'origine.

Combien d'échantillons va-t-on récolter avec cette plage du signal observé ?

Question 2

[solution n°3 p. 37]

On souhaite maintenant numériser les valeurs de ces échantillons au multiple de 5 inférieur le plus proche.

Quelles sont, dans l'ordre, les valeurs entières associées ?

Indice :

Par exemple, 0 est le multiple de 5 inférieur le plus proche de 0.3 mais aussi de 0.0.

De même, -5 est le multiple de 5 inférieur le plus proche de -2.1 et de -4.00001

Question 3

[solution n°4 p. 38]

Cette représentation discrète n'est pas très bonne car elle ne prend pas suffisamment en compte les variations des valeurs du signal.

Quels sont les deux paramètres sur lesquels on peut agir pour améliorer la discrétisation ?

VI Base numérique

Objectifs

- Découvrir la notion de base numérique ;
- Découvrir l'écriture de nombres dans une base numérique.

Mise en situation

On a l'habitude de représenter les nombres avec des chiffres de 0 à 9. C'est ce que l'on appelle le système décimal. Ce système de représentation n'est pas le seul possible, on pourrait également décider par exemple de n'utiliser que les chiffres de 0 à 5. Dans ce cas le nombre six s'écrirait 10. Les ordinateurs ne manipulent que des données binaires, c'est à dire qu'ils représentent tous les nombres avec uniquement deux chiffres : le 0 et le 1.

Si ces systèmes sont différents, ils partagent néanmoins le même système mathématique de codage : la base.

Ainsi notre système décimal est un système en base 10, et que le binaire est un système en base 2.

Base arithmétique — Az Définition

Une base arithmétique correspond au nombre de symboles, ou chiffres, utilisés pour représenter les nombres.

Base 10 (base décimale) : la base humaine — Exemple

La base décimale est la base arithmétique que l'on utilise tous les jours.

Elle utilise 10 chiffres : 0, 1, 2, 3, 4, 5, 6, 7, 8, 9.

Ici, la base est dix (10).

Représentation d'un nombre dans une base arithmétique — Méthode

Chaque nombre est une addition de puissances de sa base arithmétique. On décompose un nombre en comptant le nombre de chacune des puissances de sa base.

La puissance zéro est appelée unité.

Représenter un nombre dans la base 10

Méthode

On liste les puissances de 10 dans la base 10 jusqu'à dépasser ce nombre:

Puissance	6	5	4	3	2	1	0
Nombre	1 000 000	100 000	10 000	1000	100	10	1

Puissances de 10 en base 10

1. En partant de la plus grande puissance,
2. on note entre 0 et 9 le nombre de fois que la puissance est contenue dans le nombre,
3. et on retranche la quantité associée :
4. on obtient à la fin la représentation dans la base voulue.

Représentation de vingt-trois dans la base 10

Exemple

Si on prend le nombre vingt-trois, celui-ci contient :

- 2 fois la puissance première ($10 = 10^1$)
- 3 fois l'unité ($1 = 10^0$)

On a donc la représentation suivante de 23 dans la base 10 :

$2 \times 10^1 + 3 \times 10^0 \rightarrow 23$

Représentation de trois-mille-six-cent-cinquante-et-un dans la base 10

Exemple

Si on prend le nombre trois-mille-six-cent-cinquante-et-un, celui-ci contient :

- 3 fois la puissance troisième ($1000 = 10^3$)
- 6 fois la puissance seconde ($100 = 10^2$)
- 5 fois la puissance première ($10 = 10^1$)
- 1 fois l'unité ($1 = 10^0$)

On a donc la représentation suivante de 3651 dans la base 10 :

$3 \times 10^3 + 6 \times 10^2 + 5 \times 10^1 + 1 \times 10^0 \rightarrow 3651$

La base 60 une autre base utilisée dans l'histoire

Complément

La base 10 n'a pas été la seule base utilisée pour compter dans l'histoire. Le système sexagésimal (base 60) a aussi été utilisé par les Babyloniens et est encore utilisé aujourd'hui pour mesurer les heures et les angles. Une raison pratique de l'utilisation de cette base est que 60 est divisible par de nombreux nombres (1, 2, 3, 4, 5, 6, 10, 12, 15, 20, 30, 60).

À retenir

- On peut représenter des nombres sous plusieurs bases.
- La base que l'on utilise dans la vie de tous les jours est la base 10 mais on peut en utiliser d'autres.

Exercice : Appliquer la notion

[solution n°5 p. 38]

La représentation de vingt-trois dans la base 10 est 23.

Puissance	3	2	1	0
Nombre	1000	100	10	1

Puissances de 10 en base 10

Soit la base 5, avec les puissances suivantes :

Puissance	3	2	1	0
Nombre	125	25	5	1

Puissances de 5 en base 10

Donner la représentation de vingt-trois dans la base 5. On rappelle que la base 5 est composée des chiffres 0, 1, 2, 3 et 4.

On ne donnera que les chiffres de l'écriture de vingt-trois dans cette base.

VIII Représentation binaire

Objectifs

- Découvrir la base binaire ;
- Découvrir la représentation binaire ;
- Découvrir la notion de bit et d'octet.

Mise en situation

Un ordinateur ne sait interpréter que des séquences de symboles. Les données qu'il manipule sont donc toujours numérisées, c'est-à-dire transformées en nombre.

Et ces nombres ne sont représentés qu'avec les chiffres 0 et 1 : l'ordinateur ne manipule que du code binaire.

Il est utile de comprendre la méthode de conversion des nombres décimaux en nombres binaires. C'est ainsi que l'on verra que si on décide de représenter la lettre A par le nombre 65 (c'est la valeur numérique qui lui est associée dans le format ASCII) alors il faudra coder un A par la séquence binaire : 1 0 0 0 0 0 1.

Base binaire — Fondamental

Contrairement à la base décimale qui contient les 10 chiffres habituels, la base binaire n'en comporte que deux : 0 et 1.

Binaire — Définition

> « Le système binaire est le système de numération utilisant la base 2. On nomme couramment bit (de l'anglais *binary digit*, chiffre binaire) les chiffres de la numération binaire. Un bit peut prendre deux valeurs, notées par convention 0 et 1.
>
> Wikipédia[2] »

Pourquoi compter avec deux chiffres uniquement ? — Remarque

Fondamentalement, le processeur (l'unité qui réalise les calculs dans un ordinateur) ne traite que deux états car les éléments de base d'un processeur, les transistors, n'ont que deux états possibles : alimenté ou non-alimenté.

De même au niveau du stockage, on ne garde que des informations de présence ou d'absence d'information: ainsi on peut tout simplement n'utiliser que les deux chiffres 0 et 1.

2. https://fr.wikipedia.org/wiki/Syst%C3%A8me_binaire

Les 10 premières puissances de 2

Complément

Puissance de 2	9	8	7	6	5	4	3	2	1	0
Nombre en base 10	512	256	128	64	32	16	8	4	2	1

Puissances de 2 en base 10

Syntaxe

Pour dénoter la base d'un nombre, on l'indique en indice après le nombre.

Le nombre 7 en base 10 s'écrit 111 en base 2. Cette relation se note : $7_{10} = 111_2$

Représentation de vingt-trois dans la base binaire

Exemple

Si on prend le nombre 23_{10}, celui-ci contient :

- 1 fois la puissance quatrième de 2 ("16" en base 10)
- 0 fois la puissance troisième de 2 ("8" en base 10)
- 1 fois la puissance seconde de 2 ("4" en base 10)
- 1 fois la puissance première de 2 ("2" en base 10)
- 1 fois l'unité ("1" en base 10)

On a donc la représentation suivante 23_{10} en binaire :

$1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = (10111)_2$

Bit

Az Définition

Chaque chiffre d'une représentation binaire s'appelle un **bit**.

C'est aussi une unité de mesure de l'information que l'on note b.

Octet

Az Définition

On appelle **octet** un ensemble de huit bits.

C'est aussi une unité de mesure de l'information que l'on note B, pour *bytes*.

Un octet est capable de représenter les nombres entiers de 0 à 255.

Synonymes : on parle aussi de **mot**.

À retenir

- La base utilisée par les ordinateurs est la base 2 : soit on a de l'information, soit on a pas d'information,
- Bit et octet sont les unités utilisées pour quantifier l'information.

Exercice : Appliquer la notion

[solution n°6 p. 38]

Exercice

Quelle est le codage binaire représentant le nombre deux ? On écrira uniquement les bits.

Après avoir trouvé la réponse, on pourra la vérifier avec l'instruction Python suivante, sur Repl.it :

```
1 print("{0:b}".format(2))
```

Exercice

Quelle est le codage binaire représentant le nombre cinq ? On écrira uniquement les bits.

Après avoir trouvé la réponse, on pourra la vérifier avec l'instruction Python suivante, sur Repl.it :

```
1 print("{0:b}".format(5))
```

Exercice

On représente la lettre A par le nombre soixante-cinq dans un standard de codage des caractères nommé ASCII.

Quelle est la représentation de soixante-cinq en binaire ? On écrira uniquement les bits.

Après avoir trouvé la réponse, on pourra la vérifier avec l'instruction Python suivante, sur Repl.it :

```
1 print("{0:b}".format(65))
```

X Représentation des images bitmaps

Objectifs

- Découvrir des représentations de données d'images ;
- Découvrir la différence entre format de compression avec pertes et format de compression sans pertes.

Mise en situation

La numérisation permet de coder l'information de telle façon qu'elle soit manipulable par les ordinateurs. Les images numérisées sont ainsi affichées sur les écrans, créées par des appareils photographiques, modifiés par des logiciels de retouche ou partagées sur les réseaux.

Il existe plusieurs façons de numériser une image. Les formats bitmap que l'on utilise pour les photographies se basent sur un tableau de points (ou pixels), chacun étant associé à un nombre qui représente une couleur.

Image matricielle et codage RGB

Fondamental

La représentation la plus simple d'une image est **matricielle**. Dans cette représentation, la couleur de **chaque** pixel, élément atomique d'une image, est directement représentée dans la version numérisée de l'image.

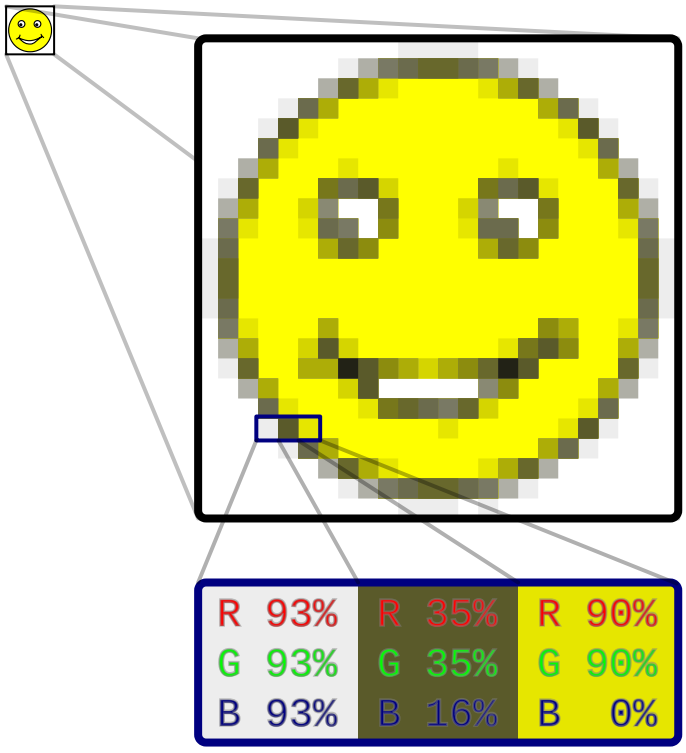


Image matricielle

Pour cela, on peut utilise le codage **RGB** : on stocke pour chaque pixel une information sur l'intensité de rouge, vert et bleu dans le pixel.

Cette intensité est comprise entre 0 et 255. Ce système de codage permet de représenter plus de 16 millions de teintes de couleurs.

Autres formats d'images

Remarque

Les formats d'images utilisés sont rarement matriciels : il existe des formats d'images qui utilisent moins d'information tout en les représentant avec une qualité qui reste très fidèle. Dans ces formats, la valeur de chaque pixel n'est pas stockée directement.

JPEG et PNG sont des formats qui utilisent des techniques avancées de traitement et compression du signal pour réduire la quantité de mémoire utilisée pour représenter une image.

Compression sans pertes et compression avec pertes

Fondamental

Après la numérisation d'un signal, on peut utiliser un algorithme de **compression sans pertes**, qui préserve intégralement l'information tout en réduisant la quantité de mémoire utilisée.

Généralement, on utilise un algorithme de **compression avec pertes**, qui dégrade l'information pour réduire davantage la quantité de mémoire utilisée.

FLAC et MP3 : des formats de compression avec et sans pertes pour l'audio

Complément

Pour l'audio, (pour *Free Lossless Audio Codec*) est un format audio de compression sans perte.

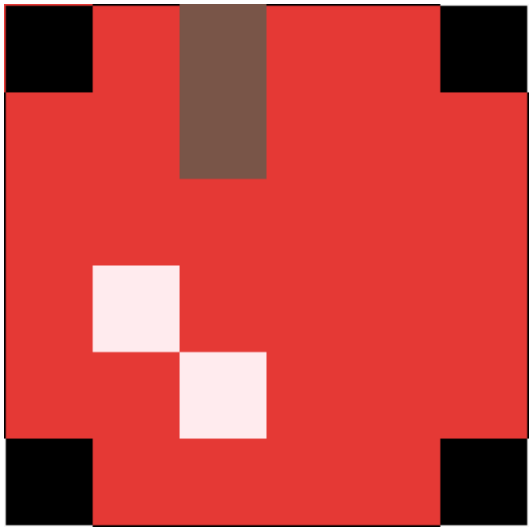
MP3 est un format audio de compression avec perte plus populaire.

Les fichiers enregistrés au format FLAC sont jusqu'à 10 fois plus volumineux que leurs semblables enregistrés au format MP3.

À retenir

- La représentation la plus simple pour les images est la représentation matricielle.
- Un système de codage populaire pour les couleurs est le format RGB.
- Il existe deux types de représentations pour les signaux numérisés : les représentations avec compression sans pertes et celles avec compression avec pertes.

XI Exercice : Appliquer la notion



On se donne l'image de 36 pixels et le codage simple suivant de couleurs élémentaires.

Binaire	Couleur
000	noir
001	blanc
010	vert
011	cyan
100	rouge
101	magenta
110	marron
111	gris

Question 1

[solution n°7 p. 39]

Représenter cette image dans un tableau de 6 lignes et 6 colonnes utilisant le codage précédent.

Question 2

[solution n°8 p. 40]

Dans le cas de cette image, pourquoi utiliser ce format plutôt que RGB ?

XII Représentation du texte

Objectif

- Découvrir des représentations de données textuelles.

Mise en situation

Lorsqu'on écrit un texte avec un ordinateur, les caractères, à l'instar des autres informations manipulées par la machine, sont représentés par des nombres. Historiquement le premier standard est ASCII qui permet de représenter 128 caractères : les lettres minuscules, majuscules, les chiffres, l'espace, des symboles comme le signe « % », et des caractères spéciaux comme le saut de ligne ou la tabulation.

ASCII est un standard américain qui n'inclut pas les caractères d'autres alphabets. Il n'y a pas les caractères accentués européens, ni les caractères arabes, ni les idéogrammes asiatiques par exemple.

Le standard Unicode, largement utilisé aujourd'hui, permet en théorie de représenter plus de 4 milliards de caractères. En pratique, il comporte aujourd'hui plus de 130.000 caractères qui permettent de traiter la quasi-totalité des langues connues.

ASCII : le premier standard pour la représentation du texte

Fondamental

Il y a vite eu une volonté de représenter les textes sous un format numérique à l'arrivée de l'informatisation. En effet, les nombres sont facilement traitables par un ordinateur. Le format ASCII, pour « *American Standard Code for Information Interchange* », a été créé à cette fin. Celui-ci utilise 7 bits pour représenter chaque caractère. Les caractères représentables en ASCII sont les lettres majuscules et minuscules, les chiffres ainsi que les symboles mathématiques et de ponctuation.

ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

C'est un choix suffisant pour de simples textes en anglais, néanmoins il ne peut pas être utilisé pour les accentuations et les caractères d'autres alphabets.

UTF-8, UTF-16 et UTF-32 : des formats d'encodage textuel modernes

Remarque

Aujourd'hui, dans des efforts d'internationalisation, des formats universels ont été mis en place pour représenter la quasi-intégralité des caractères utilisés dans les langues du monde entier.

Ces formats sont UTF-8, UTF-16 et UTF-32 qui utilisent respectivement 8, 16 et 32 bits pour représenter les caractères. Ceux-ci sont rétro-compatibles avec ASCII dans leur conception. UTF signifie *Universal Character Set Transformation Format* et ces encodages ont été développés par l'Organisation internationale de normalisation connue sous l'acronyme ISO.

La plupart des émoticônes peuvent ainsi être représentés en UTF-8.

Encodage de la lettre A en UTF-8, UTF-16 et UTF-32

Exemple

A est codé en ASCII par 41 en hexadécimal (ce qui équivaut à 65 en décimal).

UTF-8, UTF-16 et UTF-32 codent A par les 3 codes suivants :

- 41 (donc comme en ASCII)
- FEFF0041
- 00000041

À retenir

- Il existe plusieurs manières de représenter les caractères, appelées **encodages**.
- ASCII fut le premier et laisse aujourd'hui la place à UTF-8, UTF-16 et UTF-32 pour l'internationalisation.

XIII Exercice : Appliquer la notion

Le fichier `message.txt` contient un contenu que nous allons découvrir.

Enregistrer le fichier `message.txt` sur votre ordinateur
(cf. message.txt)

Question 1

[solution n°9 p. 40]

Utiliser la commande `cat` dans un terminal pour visualiser le contenu du fichier.

Indice :

```
cat message.txt
```

Question 2

[solution n°10 p. 40]

Utiliser la commande suivante dans un terminal. Quel est l'encodage du fichier ?

```
file -i message.txt
```

Question 3

[solution n°11 p. 40]

Utilisez la commande suivante pour changer le format du fichier de ISO-8859-1 à UTF-8.

```
iconv -f ISO-8859-1//TRANSLIT -t UTF-8 message.txt -o message_converti.txt
```

Vérifiez que le fichier est bien encodé en UTF-8.

Affichez à nouveau le contenu du message converti.

Indice :

```
file -i message_converti.txt
```

```
cat message_converti.txt
```

XIV Format de fichiers

Objectifs

- Découvrir la différence entre format propriétaire et format ouvert ;
- Découvrir la convention d'extension de fichiers ;
- Découvrir la structure des fichiers.

Mise en situation

Le codage en binaire n'est pas suffisant pour représenter de l'information dans un ordinateur. Il est également nécessaire de définir des formats. Un format décrit la façon de coder l'information, par exemple le fait que l'on va associer le nombre 65 à la lettre A est défini par le format ASCII.

Dans la machine, l'information est stockée sous forme de fichiers. Ces fichiers ont donc un format qui permet de définir comment les utiliser. Il existe une grande variété de formats de fichiers, par exemple :

- pour l'audio : MP3 ou FLAC,
- pour le texte : DOC ou ODT,
- pour la vidéo : MP4 ou AVI.

Format de fichier Az Définition

Un **format de fichier** définit comment l'information codée au sein d'un fichier est organisée.

Formats utilisés pour le traitement de texte Exemple

Pour le traitement de textes, il existe différents formats de fichiers. Les plus courants sont :

- TXT, qui représente le texte sans style,
- DOC, qui est le format utilisé et rendu populaire par le logiciel Word développé par Microsoft,
- ODT, qui est le format utilisé par les logiciels Open Office et Libre Office.

Formats utilisés pour les images Exemple

Pour les images, il existe une grande variété de formats. Les plus courants sont :

- PNG qui gère la transparence des images,
- JPEG qui est bien adapté pour des photographies,
- SVG qui gère les images vectorielles.

Remarque

Il faut remarquer la différence entre l'**organisation** d'un fichier et le **codage** des informations qu'il contient. Un format de fichier définit à la fois l'organisation des informations nécessaires pour représenter la ressource et leur codage.

Format ouvert et format propriétaire

Fondamental

Les formats de fichiers peuvent être **ouverts** ou **propriétaires**.

Un format ouvert se base sur un **standard public** : la manière dont est structuré le fichier est transparente. Ce format permet l'**interopérabilité** des systèmes et des logiciels.

Un format propriétaire se base sur un **standard privé** : la manière dont est structuré le fichier est uniquement connue de l'auteur.

Format de documents textes

Exemple

- Le format HTML est un format ouvert de document : n'importe quel navigateur peut afficher de tels documents et n'importe qui peut facilement modifier ce document à l'aide de n'importe quel éditeur de textes.
- Le format DWG est un format dont Autodesk est propriétaire : ce type de document ne peut être ouvert et modifié qu'avec AutoCAD, le logiciel de dessin assisté par ordinateur de texte d'Autodesk.

Extension

Fondamental

Chaque format de fichier possède **une ou plusieurs extensions** qui facilitent son identification.

Extension de fichiers multimédia

Exemple

Format	Extension
GIF : Graphics Interchange Format	.gif
JPEG	.jpeg, .jpg
AVI : Audio Video Interleave	.avi
PDF : Portable Document Format	.pdf

Images

Une extension n'est qu'une convention

⚠ Attention

L'extension du fichier facilite l'identification du type d'un fichier et aide le système d'exploitation à trouver l'application adaptée pour son utilisation.

Néanmoins, ce n'est qu'une convention : l'extension du fichier peut être changée sans affecter le contenu du fichier.

Ainsi, une image JPEG portant une extension `.txt` sera toujours une image JPEG et pourra être exploitée par tous les logiciels d'image.

Structure d'un format de données binaire

Fondamental

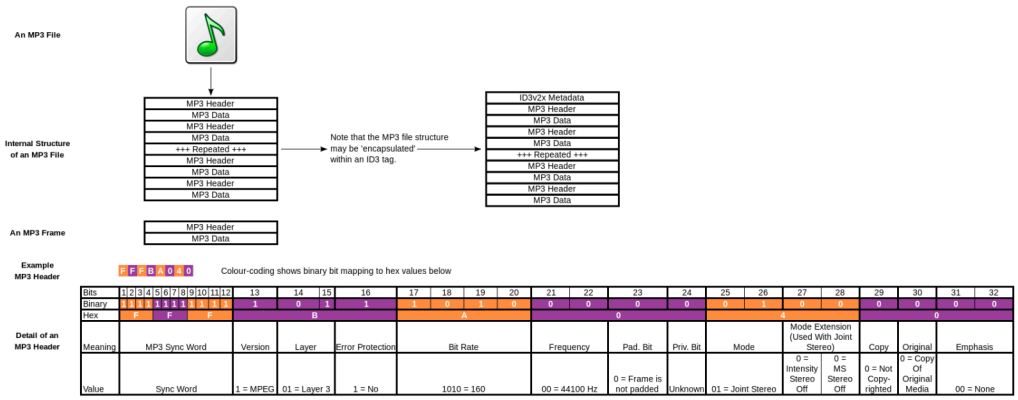
Un format de données binaire est un format qui ne stocke pas l'information sous forme de texte. Ces formats structurent les fichiers en plusieurs **segments**. Les types de segments sont les suivants :

- une **signature** qui indique le type de fichier.
- un **en-tête** qui spécifie des informations sur le contenu du fichier comme le type d'encodage, la taille de l'encodage, la version du format, la longueur du contenu, etc., ou qui contient certaines **méta-données** propres au type de fichier. On parle aussi de `header`.
- son **contenu**, tel que spécifié par le format.
- un **segment de fin de fichier** éventuellement.

Structure du format MP3

Exemple

MP3 est un fichier audio qui stocke la majorité de son information sous un format binaire. MP3 est structuré par des **datagrammes**. Chaque datagramme est composé d'un segment d'en-tête et d'un bloc de données.



Le *header* comporte toute l'information nécessaire pour pouvoir interpréter les données dans le segment suivant : on retrouve par exemple la version du format MP3 à utiliser et la fréquence d'échantillonnage.

Un fichier peut comporter un en-tête général qui spécifie des méta-données. Dans le cas de MP3, ces méta-données sont typiquement les informations sur l'enregistrement tel que l'auteur, l'album, une référence vers la pochette de l'album, etc.

À retenir

- Il existe beaucoup de formats de fichiers que l'on peut classer en deux types : les formats propriétaires et les formats ouverts.
- L'extension d'un fichier n'est qu'une convention : elle n'est là que pour aider à identifier le format.
- Les formats de fichiers complexes sont souvent structurés en segments.

XV Exercice : Appliquer la notion

On s'intéresse ici au format PNG qui est un format générique d'image très utilisé.

Pour cela on parcourt l'article Wikipédia associé : https://fr.wikipedia.org/wiki/Portable_Network_Graphics

Question 1

[solution n°12 p. 40]

Quelle est la taille minimale en octets d'un fichier ?

Indice :

Référez-vous à la section *Structure d'un fichier PNG*.

Lire maintenant la page francophone du format PNG : http://ptaff.ca/png/

Question 2

[solution n°13 p. 40]

Pourquoi le format PGN a-t-il été inventé ? Donnez deux raisons.

XVI Essentiel

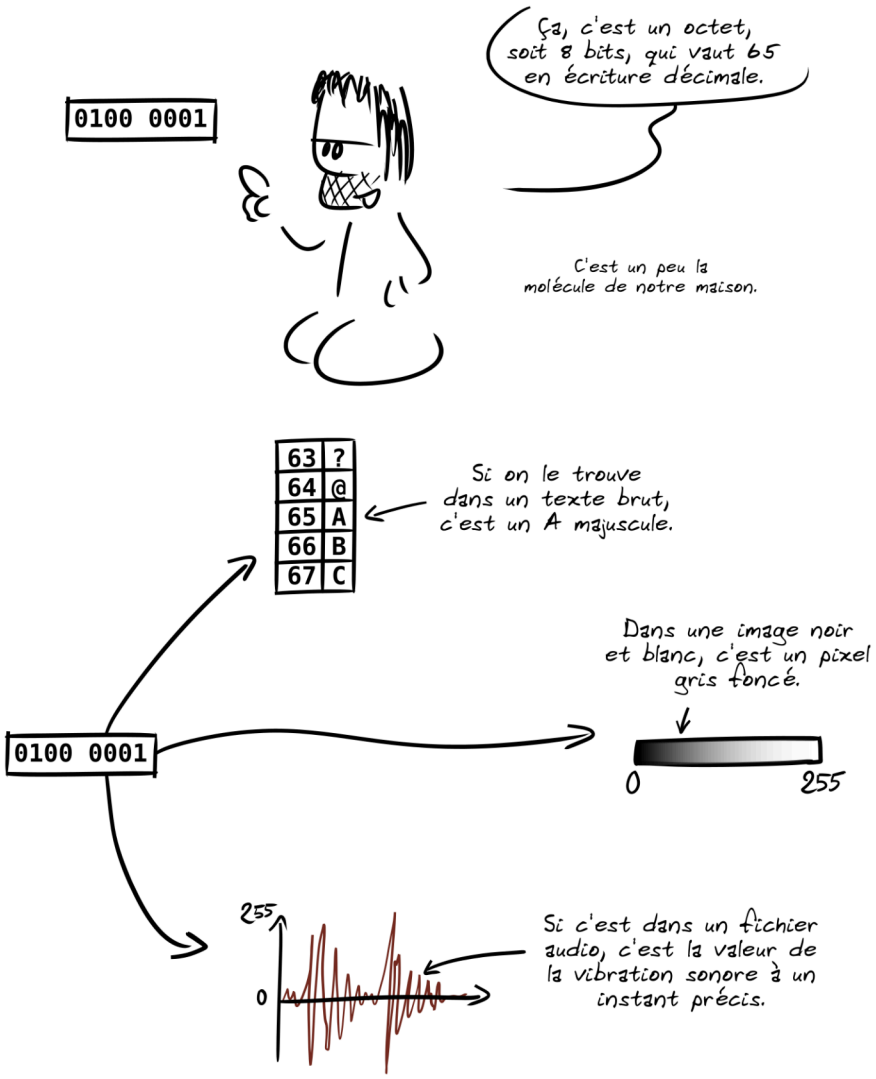
Le codage binaire est le processus qui permet de **représenter** les informations du monde qui nous entoure en séquences de 0 et de 1 que l'ordinateur peut manipuler.

Il est possible de coder en binaire n'importe quelle information. Il suffit pour cela de la **discrétiser**, c'est à dire de la découper en petits morceaux, puis de la **numériser** c'est à dire de représenter chaque morceau par un nombre.

Par exemple, on peut découper un texte selon chacun de ses caractères et associer chaque caractère existant à un nombre binaire. Nous disposons ainsi d'une méthode pour coder un **texte** en une séquence binaire.

Mais il faut également se doter d'un **format** qui permet de fixer une façon standard de coder l'information. Par exemple le standard **ASCII** pose que le Z majuscule est associé au nombre 90, qui s'écrit 101 1010 en binaire. Ainsi tous les programmes qui utilisent le format ASCII interpréteront la séquence 101 1010 comme le caractère Z.

Le même principe est appliqué pour coder les **images**, les **sons** ou les **vidéos**. On utilise simplement des méthodes de codage différentes. Par exemple pour les images bitmaps on découpe l'image en pixels et on associe chacun d'eux à un nombre qui représente sa couleur.



Bien sûr, il y a des formats plus complexes ou compressés, mais le principe reste le même :

une suite de bits, c'est une information
qu'on interprète différemment selon le contexte.

grisebouille.net/des-zeros-et-des-uns

XVII Quizz

Exercice 1 : Quiz - Culture

[solution n°14 p. 41]

Exercice

Parmi les formats suivants, quels sont les formats d'images ?

- **A** PNG
- **B** SQ
- **C** TXT
- **D** PDF
- **E** RAW
- **F** AAC

Exercice

Parmi les formats suivants, quels sont les formats ouverts ?

- **A** TXT
- **B** PDF
- **C** JPEG
- **D** 3DXML
- **E** WMA
- **F** FLAC

Exercice

Quelles sont les propriétés des images vectorielles et images matricielles ?

- **A** Les images vectorielles sont composées de flèches que l'on appelle vecteurs.
- **B** Les images matricielles sont composées de pixels.

C ☐ Les images vectorielles ont leurs propriétés de formes et de styles décrites textuellement.

D L'information portée par les images matricielles est essentiellement numérique.

E Les images vectorielles contiennent aussi des pixels.

F ☐ Les images matricielles ont une palette de couleurs bien plus grande que les images vectorielles.

Exercice

Parmi les appareils ci-dessous, lesquels contiennent des convertisseurs analogique numérique ?

A Appareil photo numérique

B Une vieille radio

C Smartphone

D Four micro-onde

E Télévision

F Lecteur MP3

G Liseuse

H Scanner de documents

Exercice 6 : Quiz - Méthode

[solution n°15 p. 42]

Exercice

Vous disposez de nombreuses photos de vacances sous formes d'images matricielles dans un format de compression sans pertes. Vous voulez stocker ces images sur un disque dur vierge, cependant celui-ci ne contient pas assez d'espace mémoire. Vous voulez tout de même mettre les photos sur celui-ci en gardant une bonne qualité d'image ; que réalisez-vous ?

A ○ Vous les convertissez sous un format d'image vectorielle car ce type de format permet d'avoir des fichiers plus légers.

- **B** Vous réduisez leur résolution pour réduire leur taille mémoire.
- **C** Vous utilisez un format d'image avec compression avec pertes.

Exercice

Vous venez d'inventer un format de compression révolutionnaire et voulez lancer une entreprise sur celui-ci. Vous avez fait le choix d'appeler votre entreprise "Pied Piper" mais vous n'avez pas encore choisi le type de format à utiliser. En tant que jeune entrepreneur, vous voulez garder la parenté et la maîtrise entière sur ce format afin d'avoir et maintenir un avantage concurrentiel. Quel est le mieux pour vous ?

- **A** Distribuer ce format de compression sous une forme propriétaire.
- **B** Distribuer ce format de compression sous un standard ouvert.

Exercice

Vous êtes une équipe de scientifiques en biologie. Vous réalisez des expériences et vous voulez en publier les résultats pour la communauté à des fins d'accessibilité et de reproductibilité. Cependant, il n'existe vraiment pas de format de fichier disponible pour votre type de données. Quel est le mieux pour vous ?

- **A** Distribuer ce format de compression sous une forme propriétaire.
- **B** Distribuer ce format de compression sous un standard ouvert.

Exercice 10 : Quiz - Structure fichier

[solution n°16 p. 44]

Exercice

Voici un extrait issu d'un fichier PCB qui décrit l'agencement des atomes d'une protéine.

```
HEADER    EXTRACELLULAR MATRIX                    22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...
ATOM      1  N   PRO A   1       8.316  21.206  21.530  1.00 17.44           N
ATOM      2  CA  PRO A   1       7.608  20.729  20.336  1.00 17.44           C
ATOM      3  C   PRO A   1       8.487  20.707  19.092  1.00 17.44           C
ATOM      4  O   PRO A   1       9.466  21.457  19.005  1.00 17.44           O
```

```
ATOM      5  CB  PRO A   1       6.460  21.723  20.211  1.00 22.26           C
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
HETATM  132  OXT ACY   401       4.306  23.101  12.291  1.00 21.19           O
```

Parmi les propositions suivantes, quels sont les mots clefs en début de ligne associés aux méta-données (date, titre, commentaire, etc.) du fichier ?

- **A** AUTHOR
- **B** HEADER
- **C** ATOM
- **D** PRO
- **E** HETATM
- **F** REMARK
- **G** EXPDTA

Exercice

À quel format de fichier correspond le contenu ci-dessous ?

```svg
<svg height="100" width="100">
  <circle cx="50" cy="50" r="40" stroke="black" stroke-width="3" fill="red" />
</svg>
```

- **A** SVG
- **B** TXT
- **C** PDF
- **D** MP3

Exercice

On dispose d'un fichier sans extension et dont on ne connaît pas le format dont voici les premières lignes.

```
%PDF-1.4
%äüöß
2 0 obj
<</Length 3 0 R/Filter/FlateDecode>>
stream
x�=��
?1 E����v?���0??�~��?�
```

Quelles informations a-t-on sur le fichier rien qu'avec cet extrait ?

- **A** C'est un fichier de format PDF.
- **B** Il contient une signature.
- **C** Il contient un en-tête.
- **D** Il contient des données binaires.
- **E** On ne pourra jamais récupérer le contenu du fichier puisque l'on ne peut pas modifier l'extension.

Solutions des exercices

Solution n°1

[exercice p. 6]

Selon l'apocryphe de Camus :

« Mal nommer les choses c'est ajouter au malheur du monde. »

Dans la langue française, on utilise des concepts que l'on représente par des mots pour s'exprimer. On dispose de plusieurs codages avec chacun leurs symboles. Mettez en correspondance ces symboles et ces codages.

On peut écrire les mots : on dispose donc d'un codage écrit composé de symboles qui sont	**On peut dire les mots : on dispose d'un codage vocal composé de symboles qui sont**	**On peut mimer les mots : on dispose d'un codage — la langue des signes — composé de symboles qui sont**
Des lettres	Des phonèmes	Des gestes

Solution n°2

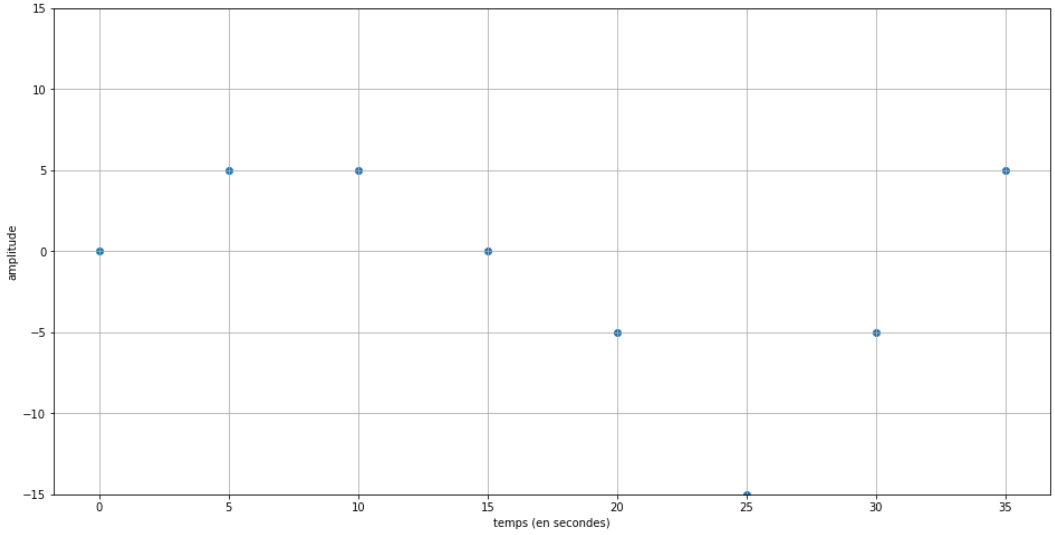
[exercice p. 10]

On obtient 8 valeurs du signal aux temps : 0, 5, 10, 15, 20, 25, 30, 35.

Solution n°3

[exercice p. 10]

On obtient après la discrétisation et la numérisation le signal suivant :



Signal analogique

On obtient les valeurs suivantes : 0, 5, 5, 0, -5, - 15, -5, 5.

Solution n°4

[exercice p. 10]

On peut réduire la période temps (appelée période d'échantillonnage) de 5 secondes à par exemple 1 seconde. En augmentant le nombre de données échantillonnées, on se rapproche de plus en plus du signal initial.

On peut également augmenter la **résolution** du signal échantillonné, par exemple en numérisant les valeurs à l'entier inférieur le plus proche. Ainsi on se rapproche de plus en plus des valeurs échantillonnées.

Solution n°5

[exercice p. 14]

La représentation de vingt-trois dans la base 10 est 23.

Puissance	3	2	1	0
Nombre	1000	100	10	1

Puissances de 10 en base 10

Soit la base 5, avec les puissances suivantes :

Puissance	3	2	1	0
Nombre	125	25	5	1

Puissances de 5 en base 10

Donner la représentation de vingt-trois dans la base 5. On rappelle que la base 5 est composée des chiffres 0, 1, 2, 3 et 4.

On ne donnera que les chiffres de l'écriture de vingt-trois dans cette base.

43

Vingt-trois contient 4 fois la puissance première (5) et 3 fois 1, donc s'écrit (43)_5.

Solution n°6

[exercice p. 18]

Exercice

Quelle est le codage binaire représentant le nombre deux ? On écrira uniquement les bits.

Après avoir trouvé la réponse, on pourra la vérifier avec l'instruction Python suivante, sur Repl.it :

```
1 print("{0:b}".format(2))
```

10

Deux est par définition la puissance première de 2.

La représentation de 2_{10} est donc 10_2.

Exercice

Quelle est le codage binaire représentant le nombre cinq ? On écrira uniquement les bits.

Après avoir trouvé la réponse, on pourra la vérifier avec l'instruction Python suivante, sur Repl.it :

```
1 print("{0:b}".format(5))
```

101

Cinq contient :

- 1 fois la puissance seconde de 2 ("4" en base 10)
- 0 fois la puissance première de 2 ("2" en base 10)
- 1 fois l'unité ("1" en base 10)

On a donc la représentation suivante de cinq dans la base 2 : 101_2

Exercice

On représente la lettre A par le nombre soixante-cinq dans un standard de codage des caractères nommé ASCII.

Quelle est la représentation de soixante-cinq en binaire ? On écrira uniquement les bits.

Après avoir trouvé la réponse, on pourra la vérifier avec l'instruction Python suivante, sur Repl.it :

```
1 print("{0:b}".format(65))
```

1000001

Soixante-cinq contient :

- la puissance sixième ($2^6 = 64$)
- l'unité (1)

On a donc la représentation suivante de soixante-cinq dans la base 2 : 1000001_2

Solution n°7

[exercice p. 21]

000	100	110	100	100	000
100	100	110	100	100	100
100	100	100	100	100	100
100	001	100	100	100	100
100	100	001	100	100	100
000	100	100	100	100	000

Solution n°8

[exercice p. 21]

Ce codage de couleur nécessite tout simplement beaucoup moins de bits pour représenter les couleurs.

Il faut dans ce format 3 bits pour représenter toutes les couleurs de l'image

Dans le cas de RGB il faut 3 octets, soit 3 × 8 = 24 bits, c'est à dire 8 fois plus de bits.

Solution n°9

[exercice p. 24]

```
1 Ceci est une solution  Ã©lÃ©gante !
```

Solution n°10

[exercice p. 24]

```
1 message.txt: text/plain; charset=iso-8859-1
```

L'encodage du fichier est iso-8859-1.

Solution n°11

[exercice p. 24]

Résultat de : file -i message_converti.txt

```
1 message_converti.txt: text/plain; charset=utf-8
```

Résultat de : cat message_converti.txt

```
1 Ceci est une solution  élégante !
```

On dit que le premier fichier était **mal encodé**. En effet, le format ISO-8859-1, ou plus communément appelé `latin1`, ne permet pas de représenter les caractères accentués. Ceux-ci s'affichent mal.

Solution n°12

[exercice p. 29]

On a un fichier de taille minimale s'il n'a pas de contenu. Dans le cas de PNG, si le fichier ne contient pas d'information, le chunk IDAT est vide.

En additionnant la taille des autres chunks, on trouve une taille minimale de :

8 + 15 + 12 = 35 octets.

Solution n°13

[exercice p. 29]

Une première raison de la création de PNG était le remplacement des images aux format GIF qui était propriétaire et qui ne permettait pas une utilisation libre.

Une seconde raison de la création de PNG était la nécessité d'un format d'images aux spécifications plus simples et aux capacités techniques améliorées.

Solution n°14

[exercice p. 32]

Exercice

Parmi les formats suivants, quels sont les formats d'images ?

- [x] **A** PNG
- [ ] **B** SQ C'est un format d'archive.
- [ ] **C** TXT C'est un format de fichiers texte.
- [ ] **D** PDF C'est un format de document.
- [x] **E** RAW Il s'agit du format brut d'images issues d'appareil photo.
- [ ] **F** AAC C'est un format audio.

Exercice

Parmi les formats suivants, quels sont les formats ouverts ?

- [x] **A** TXT
- [x] **B** PDF PDF est un exemple de format propriétaire qui a été ouvert et standardisé en 2008.
- [x] **C** JPEG
- [ ] **D** 3DXML C'est un format propriétaire de Dassault Systèmes pour son logiciel Catia
- [ ] **E** WMA C'est un format propriétaire de Microsoft.
- [x] **F** FLAC

Exercice

Quelles sont les propriétés des images vectorielles et images matricielles ?

- [x] **A** Les images vectorielles sont composées de flèches que l'on appelle vecteurs.
- [x] **B** Les images matricielles sont composées de pixels.
- [ ] **C** Les images vectorielles ont leurs propriétés de formes et de styles décrites textuellement.

D L'information portée par les images matricielles est essentiellement numérique.

E Les images vectorielles contiennent aussi des pixels.

F Les images matricielles ont une palette de couleurs bien plus grande que les images vectorielles.

Exercice

Parmi les appareils ci-dessous, lesquels contiennent des convertisseurs analogique numérique ?

A Appareil photo numérique

B Une vieille radio — Une radio convertit une onde radio en une onde sonore sans numériser de signal.

C ☐ Smartphone — Un smartphone dispose de plusieurs convertisseurs analogiques numériques.

D Four micro-onde

E ☐ Télévision — Une télévision reçoit une onde infrarouge à partir de la télécommande et dispose d'un convertisseur pour numériser le signal.

F Lecteur MP3

G Liseuse

H Scanner de documents

Solution n°15

[exercice p. 33]

Exercice

Vous disposez de nombreuses photos de vacances sous formes d'images matricielles dans un format de compression sans pertes. Vous voulez stocker ces images sur un disque dur vierge, cependant celui-ci ne contient pas assez d'espace mémoire. Vous voulez tout de même mettre les photos sur celui-ci en gardant une bonne qualité d'image ; que réalisez-vous ?

A

Vous les convertissez sous un format d'image vectorielle car ce type de format permet d'avoir des fichiers plus légers.

Il n'est souvent pas possible de réaliser la conversion d'images matricielles vers des images vectorielles : cela est particulièrement vrai dans le cas de photos.

B

Vous réduisez leur résolution pour réduire leur taille mémoire.

C'est une solution qui fonctionne mais elle est brutale : cela réduit drastiquement la taille des fichiers mais vous perdez directement en qualité d'images.

C

Vous utilisez un format d'image avec compression avec pertes.

C'est la meilleure solution : vous pouvez gagner beaucoup en espace mémoire tout en gardant une qualité très bonne d'image.

Exercice

Vous venez d'inventer un format de compression révolutionnaire et voulez lancer une entreprise sur celui-ci. Vous avez fait le choix d'appeler votre entreprise "Pied Piper" mais vous n'avez pas encore choisi le type de format à utiliser. En tant que jeune entrepreneur, vous voulez garder la parenté et la maîtrise entière sur ce format afin d'avoir et maintenir un avantage concurrentiel. Quel est le mieux pour vous ?

A Distribuer ce format de compression sous une forme propriétaire.

B Distribuer ce format de compression sous un standard ouvert.

Le format propriétaire est le plus adapté ici : il vous permet de rester entièrement maître de votre création.

Exercice

Vous êtes une équipe de scientifiques en biologie. Vous réalisez des expériences et vous voulez en publier les résultats pour la communauté à des fins d'accessibilité et de reproductibilité. Cependant, il n'existe vraiment pas de format de fichier disponible pour votre type de données. Quel est le mieux pour vous ?

A Distribuer ce format de compression sous une forme propriétaire.

B Distribuer ce format de compression sous un standard ouvert.

Un format reposant sur un standard ouvert est le plus adapté dans cette situation : il permet d'avoir une représentation des données facilement exploitable, favorise la collaboration et la transparence.

C'est par exemple historiquement ce qui s'est passé avec le format PDB (*Protein Data Bank*) pour la représentation de protéines. Plus d'informations ici : http://www.wwpdb.org/documentation/file-format-content/format33/sect1.html

Solution n°16

[exercice p. 34]

Exercice

Voici un extrait issu d'un fichier PCB qui décrit l'agencement des atomes d'une protéine.

```
HEADER    EXTRACELLULAR MATRIX                    22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...
ATOM      1  N   PRO A   1       8.316  21.206  21.530  1.00 17.44           N
ATOM      2  CA  PRO A   1       7.608  20.729  20.336  1.00 17.44           C
ATOM      3  C   PRO A   1       8.487  20.707  19.092  1.00 17.44           C
ATOM      4  O   PRO A   1       9.466  21.457  19.005  1.00 17.44           O
ATOM      5  CB  PRO A   1       6.460  21.723  20.211  1.00 22.26           C
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
HETATM  132  OXT ACY   401       4.306  23.101  12.291  1.00 21.19           O
```

Parmi les propositions suivantes, quels sont les mots clefs en début de ligne associés aux méta-données (date, titre, commentaire, etc.) du fichier ?

A AUTHOR

B HEADER

C ATOM

D PRO

E HETATM

F REMARK

G EXPDTA

Les mots clefs HEADER, TITLE, AUTHOR et EXPDTA définissent les informations sur les méta-données du fichier, dans le cas ici : nom de la molécule, date, auteur, information sur l'expérience.

Exercice

À quel format de fichier correspond le contenu ci-dessous ?

```svg
<svg height="100" width="100">
  <circle cx="50" cy="50" r="40" stroke="black" stroke-width="3" fill="red" />
</svg>
```

A SVG

B TXT

C PDF

D MP3

C'est un fichier SVG : on voit simplement cela avec la balise <svg> par exemple.

Exercice

On dispose d'un fichier sans extension et dont on ne connaît pas le format dont voici les premières lignes.

```
%PDF-1.4
%äüöß
2 0 obj
<</Length 3 0 R/Filter/FlateDecode>>
stream
x@=@@
?1 E@@@@v?@@@0??@~@@?@
```

Quelles informations a-t-on sur le fichier rien qu'avec cet extrait ?

A C'est un fichier de format PDF.

B Il contient une signature.

C Il contient un en-tête.

D Il contient des données binaires.

E On ne pourra jamais récupérer le contenu du fichier puisque l'on ne peut pas modifier l'extension.

Il y a une signature en haut du fichier qui indique que c'est un document PDF.

De plus il dispose d'un petit en-tête qui indique les informations sur le format du contenu.

Il contient aussi des données binaires qui ne sont pas interprétables directement si on l'ouvre avec un éditeur de texte.

Enfin, on peut renommer le fichier avec l'extension correcte `.pdf` et ensuite l'utiliser.

Crédits des ressources

p. 3

Attribution - Partage dans les Mêmes Conditions - stph à partir de dessins de Gee[3] *(framalab.org/gknd-creator)*

Codeur rotatif ROD 420 p. 9

Attribution - Partage dans les Mêmes Conditions - Victor Korniyenko

Image matricielle p. 19

Universel - Transfert dans le Domaine Public - Gringer, fr.wikipedia.org/wiki/Image_matricielle

grisebouille.net/des-zeros-et-des-uns p. 31

Attribution - Partage dans les Mêmes Conditions - Gee

3. https://grisebouille.net/