

## Chapitre 3 - Données structurées

### I - Un peu d'histoire

Date	Support	Capacité	Exemple d'utilisation
1801	Cartes perforées	Quelques octets	Programmation des métiers à tisser Jacquard
1890	Cartes perforées d'Hollerith	80 octets	Recensement américain
1928	Ruban magnétique	Quelques Ko par mètre	Enregistrement audio, premiers calculs
1932	Tambour magnétique	Environ 10 Ko	Premiers ordinateurs (IBM 650)
1956	Disque dur IBM 305 RAMAC	5 Mo	Stockage dans les entreprises
1963	Cassette magnétique	Quelques centaines de Ko	Mini-ordinateurs, enregistrement audio
1967	Disque flexible (8 pouces)	80 Ko	Chargement de systèmes d'exploitation
1976	Disquette 5 pouces 1/4	110 Ko à 1,2 Mo	Ordinateurs personnels (Apple II, IBM PC)
1981	Disquette 3 pouces 1/2	720 Ko à 1,44 Mo	Stockage portable pour ordinateurs personnels
1979	Disque compact (CD)	700 Mo	Musique, logiciels, données personnelles

<b>1984</b>	Disque dur IDE	20 Mo à 100 Mo	Stockage local sur ordinateurs personnels
<b>1995</b>	DVD	4,7 Go	Films, logiciels, données volumineuses
<b>2000</b>	Clé USB	8 Mo à plusieurs Go	Transport et partage de fichiers
<b>2006</b>	Blu-ray	25 Go à 50 Go	Films HD, jeux vidéo
<b>2010</b>	Disques SSD	128 Go à plusieurs To	Stockage rapide pour ordinateurs et serveurs
<b>2013</b>	Stockage en nuage (Cloud)	Virtuellement illimitée	Sauvegarde, synchronisation, collaboration
<b>2022</b>	Stockage ADN expérimental	Plusieurs pétaoctets par gramme	Archivage à long terme
<b>Futur</b>	Stockage quantique	Capacité promettant d'être massive	Applications scientifiques et données massives

Unité	Symbole	Valeur	Puissance de 10 équivalente
Bit	b	1 bit	$10^0$
Octet	o ou B	8 bits	$10^0$
Kiloctet	Ko	1 Ko = 1 000 o	$10^3$
Mégoctet	Mo	1 Mo = 1 000 Ko	$10^6$
Gigaoctet	Go	1 Go = 1 000 Mo	$10^9$
Téraoctet	To	1 To = 1 000 Go	$10^{12}$
Pétaoctet	Po	1 Po = 1 000 To	$10^{15}$
Exaoctet	Eo	1 Eo = 1 000 Po	$10^{18}$

 Une musique 4 Mo	 Une photo 6 Mo	 Un document 50 Ko	 Un film 700 Mo
 Un ordinateur récent de 500 Go à 4 To	 Une clé USB / carte mémoire de 8 Go à 200 Go		 Une disquette 1.4 Mo

## II - Quelques définitions

### Définition 1

#### Notion de donnée

Une donnée (data en anglais) est une valeur décrivant un objet, une personne , un événement digne d'intérêt pour celui qui choisit de la conserver.

Par exemple, le numéro de téléphone d'un contact peut être une donnée .

### Définition 2

#### Descripteur

Un descripteur permet de décrire un objet.

Un objet peut posséder plusieurs descripteurs ( par exemple, des descripteurs permettant de caractériser un objet de type contact : nom , prénom, numéro de téléphone)

### Définition 3

#### Collection

Une collection regroupe des objets partageant les mêmes descripteurs ( par exemple les contacts d'un carnet d'adresse).

Une structure de table est alors utile pour représenter une collection : on y place les objets en ligne et les descripteurs en colonne.

Collection				
Descripteurs	Nom	Capitale	Hymne	Superficie (km <sup>2</sup> )
Une valeur du descripteur « Nom »	France	Paris	La Marseillaise	632 734
	Chine	Pékin	La Marche des volontaires	9 596 961
	États-Unis	Washington	The Star-Spangled Banner	9 833 517
Un objet	Argentine	Buenos Aires	Himno Nacional Argentino	2 791 810

### Définition 4

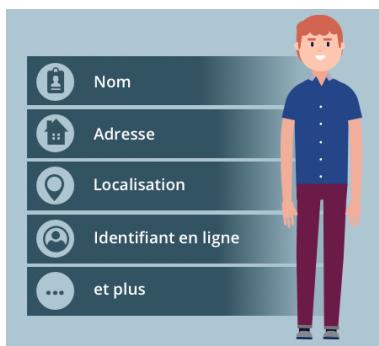
#### Donnée personnelle

Une donnée est qualifiée de **donnée personnelle** si elle se rapporte à une **personne identifiée ou identifiable** ( Art 4 RGPD ).

Exemples de données personnelles :

- nom et/ou prénom
- adresse IP
- une photographie
- un numéro de téléphone
- une donnée biométrique
- des données de localisation (comme latitude et longitude)

Le RGPD (voir ci dessous) est alors là pour encadrer l'utilisation de ces données .



## Définition 5

### Données ouvertes ou Open data

Les données ouvertes ou open data sont des données numériques dont **l'accès et l'usage sont laissés libres aux usagers**. Elles peuvent être d'origine publique ou privée, produite notamment par une collectivité, un service public ou une entreprise.

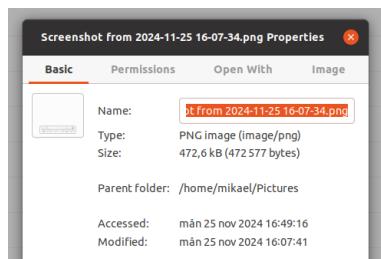
## Définition 6

### Métadonnées

A tout fichier informatique sont associées des métadonnées qui permettent d'en décrire le contenu .

Ces métadonnées varient selon le type de fichier. Il peut s'agir de la date et des coordonnées de géolocalisation d'une photographie, de l'auteur et du titre d'un fichier texte , etc...

On accède aux métadonnées d'un fichier personnel par un clic droit sur le nom de fichier.



## III - Les formats des fichiers

### 1) Quelques formats de fichiers courants



## 2) Formats de fichiers pour l'open data

### Format CSV

Dans un fichier au format CSV (**Comma Separated Values**) , les données sont présentés dans un fichier texte, les valeurs étant séparés par un caractère spécifique.

Les caractères les plus connus comme séparateurs sont la virgule ou le point-virgule.

Ce format est très facile à générer et à manipuler. Chaque ligne du fichier CSV correspond à une ligne du tableau et chaque valeur séparée par une virgule correspond à une colonne du tableau.

#### Exemple :

On écrit les données dans un fichier CSV à l'aide d'un traitement de texte.

```
Nom;Prénom;surnom;mot de passe;ville  
PHILIPPE;frederic;dédé;45§7;villereau  
PILLOT;Jean;jannot;@4r3e;lyon  
HENRY;edouard;doudou;$456;Lille
```

Ce fichier de donnée au format .csv peut être importé dans un logiciel tableur (Excel, Libreoffice Calc...) :

	A	B	C	D	E
1	Nom	Prénom	surnom	mot de passe	ville
2	PHILIPPE	frederic	dédé	45§7	villereau
3	PILLOT	Jean	jannot	@4r3e	lyon
4	HENRY	edouard	doudou	\$456	Lille
5					
6					

### Format JSON

Dans un fichier au format JSON (**JavaScript Object Notation**) , les données sont présentées dans un fichier texte en utilisant une syntaxe proche d'un langage de programmation très utilisé sur internet : le JavaScript.

Ce format a pour intérêt de stocker des données plus complexes que celles présentes dans un format CSV. Ce format permet de décrire un objet en définissant des paires descripteur/valeur séparées par le caractère « : » et chaque paire est séparée par le caractère « , ».

La définition d'un objet commence par une accolade ouvrante et se termine par une accolade fermante. On peut également définir un descripteur comme un objet.

#### Exemple :

On écrit les données dans un fichier JSON à l'aide d'un traitement de texte.

Dans cet exemple, on définit un objet avec 4 descripteurs (Espèce, age, race, particularité). Le descripteur "particularité" est également un objet possédant deux descripteurs (couleur yeux et couleur pelage).

```
{  
    "espèce": "chien",  
    "age": "6 ans",  
    "race": "cocker",  
    "particularité": {  
        "couleur yeux": " marron",  
        "couleur pelage": "noir et blanc"  
    }  
}
```

Ce fichier de donnée au format .json peut être affiché dans une page HTML (page web) en utilisant un programme informatique (écrit en Python par exemple) :

```
espèce:          "chien"  
age:             "6 ans"  
race:            "cocker"  
particularité:  
    couleur yeux: " marron"  
    couleur pelage: "noir et blanc"
```

## IV - Impacts sur les pratiques humaines

### 1) Les données personnelles et le RGPD

#### Définition 7

Le règlement général de protection des données (RGPD) est un texte réglementaire européen qui encadre le traitement des données de manière égalitaire sur tout le territoire de l'Union européenne (UE). Il est entré en application le 25 mai 2018.

Le RGPD s'inscrit dans la continuité de la loi française « Informatique et Libertés » de 1978, modifiée par la loi du 20 juin 2018 relative à la protection des données personnelles, établissant des règles sur la collecte et l'utilisation des données sur le territoire français. Il a été conçu autour de trois objectifs :

- renforcer les droits des personnes
- responsabiliser les acteurs traitant des données
- crédibiliser la régulation grâce à une coopération renforcée entre les autorités de protection des données.



#### Les six grands principes du RGPD

1. Ne collectez que les données vraiment nécessaires pour atteindre votre objectif.
2. Soyez transparent.
3. Organisez et facilitez l'exercice des droits des personnes.
4. Fixez des durées de conservation.
5. Sécurisez les données et identifiez les risques.
6. Inscrivez la mise en conformité dans une démarche continue.

## 2) Le BIG DATA

## Définition 8

Le **Big Data** est un terme utilisé pour décrire l'abondance des données numériques et l'émergence de moyens développés pour y accéder et l'analyser.

Son rôle est de traiter des informations pour acquérir de nouvelles connaissances . Pour en extraire du sens, il faut trier d'énormes volumes de données.

Aujourd’hui le Big Data est utilisé pour apprendre et résoudre des problèmes dans de nombreuses disciplines notamment grâce aux technologies d’intelligence artificielle qui reposent dessus.

En 2024, le volume des données générées quotidiennement dépasse les 4 trillions d'octets, provenant de multiples sources. Notamment les messages envoyés, vidéos publiées, informations climatiques, signaux GPS, enregistrements transactionnels d'achats en ligne et bien d'autres encore.



## Quelques chiffres clés du BIG DATA en 2024

- Environ **28,77 millions de téraoctets (0,33 zettaoctets)** de données sont créés chaque jour, soit **120 zettaoctets** par an. Ce volume est alimenté par l'essor des appareils connectés et l'expansion de l'activité numérique mondiale.
  - Environ **97 %** des entreprises ont investi dans des solutions Big Data, mais seulement 24 % utilisent pleinement ces données pour prendre des décisions stratégiques. Par ailleurs, 40 % des entreprises utilisent le Big Data pour mieux comprendre leurs clients et leurs comportements d'achat.
  - On recense actuellement environ **8 100 datacenters** dans le monde, dont près d'un tiers se trouvent aux États-Unis (environ 2 700). L'Allemagne (466), le Royaume-Uni (449), et la Chine (415) suivent. La France se positionne 8ème avec environ 247 datacenters.

### 3) Le cloud

## Définition 9

Le **cloud** désigne l'ensemble des ressources informatiques (stockage, services) disponibles sur internet plutôt que localement sur un ordinateur.

En utilisant la messagerie (webmail) tel que Gmail, Hotmail ou Yahoo, on utilise sans nous en rendre compte un service dans le cloud.

De la même façon, si on utilise un service de stockage tel que Dropbox ou Google Drive, on utilise des services du cloud qui utilisent la puissance de nombreux serveurs informatiques mutualisés distants, plutôt que de stocker les fichiers sur notre propre ordinateur.

## 4) Impacts environnementaux

Mais où sont stockés nos fichiers s'ils ne sont plus sur nos ordinateurs ?

Ils sont dans des **Data Centers**.

Il s'agit d'endroit physique qui possède de nombreux serveurs pour répondre aux besoins de plus en plus croissant de stockage.

Les Data Centers ne sont pas déterminés par leur taille physique. Les petites entreprises peuvent utiliser une petite salle où sont juxtaposés plusieurs serveurs et espaces de stockage interconnectés. Les entreprises informatiques de grande envergure, comme Facebook, Amazon ou Google, peuvent quant à elles remplir un immense entrepôt.



De telles installations dégagent cependant énormément de chaleur et doivent être refroidies pour éviter toute panne, ce qui induit une consommation électrique très élevée.

- En 2024, on estime qu'environ 20 % de la consommation électrique liée au numérique provient des data centers.
- En 2024, on estime que les datacenters consomment environ 2 à 4 % de l'électricité mondiale.

## V - Exercices

### Exercice 1 Les données dans notre société

Les données sont devenues un enjeu pour notre société ; présentes dans de nombreux domaines ( santé, l'éducation, l'industrie, la sécurité... ), on a ainsi vu apparaître de nouveaux termes comme : Big Data, Open Data... De nouveaux métiers sont créés : Architecte Big Data, Data scientist, ... de nouvelles technologies sont développées : le cloud computing...

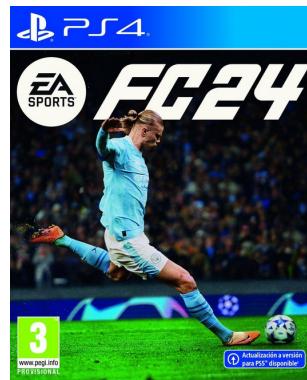
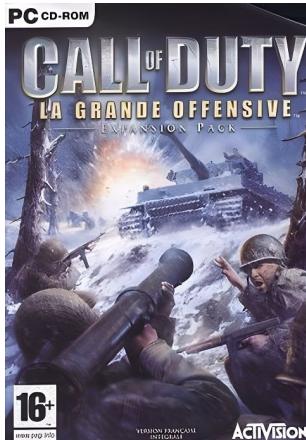
L'utilisation et la maîtrise du big data suscite beaucoup d'enthousiasme mais également des inquiétudes en particulier sur la protection des données personnelles.

En consultant le cours que vous avez à votre disposition, répondez aux questions suivantes :

- Etablir une définition des mots suivants : big data , open data , métadonnées , données ouvertes, données personnelles .
- Qu'est-ce que le RGPD ? De quand date-t-il ?
- Que veut dire selon vous le terme "Data scientist" ?

### Exercice 2 Collection, objet, descripteur, valeur

Choisir 4 descripteurs permettant de classer dans un tableau cette collection de jeux vidéo, puis compléter le tableau.



### Exercice 3 Format des fichiers

Pour assurer leur persistance et leurs échanges , les données sont structurées sous différents formats c'est à dire un mode d'organisation qui les rendent lisibles, faciles à mettre à jour.

Ainsi, quand on enregistre un fichier, une extension accompagne ce fichier.

Exemple : L'extension « .pdf » signale à un ordinateur le type de fichier auquel on a affaire en l'occurrence une fichier texte.

Reproduire le tableau ci-dessous et compléter le avec les formats de fichiers de la section III de votre cours.

Type de fichier	Texte	Image	Vidéo	Musique	Web	Données tabulaires	Autres
Extension	.pdf					.CSV	

## **Exercice 4 Parcours PIX**

Connectes-toi sur PIX et entre le code suivant : **ALXPEJ626** pour commencer les exercices sur les données structurées.

## **Exercice 5 Artificial intelligence (par groupe de 2 élèves, EVALUÉ)**

### **PART 1 - Artificial intelligence explained in 2 minutes**

Download the video "Artificial intelligence explained in 2 minutes.mp4" on EcoleDirecte, watch it and answer the following questions.

- Question 1 : Give a short definition of what is AI ?
- Question 2 : What is the first way to achieve AI ?
- Question 3 : What is the second way to achieve AI ? How is this second way called ?
- Question 4 : Give 6 examples of AI described in this video.

### **PART 2 - Extracts of an article from BBC Science Focus magazine**

#### **Are there different kinds of AI ?**

Mention AI, and most people think of 'deep learning'. This kind of AI is loosely inspired by the way our brains work. It uses lots of computers to simulate large networks of artificial 'neurons', which are then trained, typically using humongous amounts of data, until they've learned to do what we want them to – for example, understanding speech.

This training is the slow and resource-heavy part. Once trained, even a phone can then run the AI and instantly perform the right function, such as obeying your voice command. Deep learning is just one kind of AI, among thousands of others.

- Question 5 : This article gives the 3 main components of AI. What are they ?
- Question 6 : The notion of training if mainly used in this article. Explain what is the meaning of training.

#### **What else is AI used for ?**

AI has become a ubiquitous technology. When you unlock your phone by looking at it, an AI has recognised your face. When you speak to your TV or smart speaker, an AI has recognised your voice. When you take a photo with your phone or digital camera, an AI identifies elements in the foreground to help blur the background, and combines several photos taken with different exposures to construct a perfect picture.

- Question 7 : What does this sentence mean : "AI has become a ubiquitous technology" ?

#### **Can we trust AI ?**

There are always downsides to technologies. If an AI is trusted too much, then we may get ourselves into trouble – this is why driverless cars will always need a human override option.

If we train AI with biased data then the AI will also be biased, as studies have shown where AIs recognise white male faces better than others. Some worry that AI will lead to job losses, which may be true, but AI will also create many jobs. AI is nothing new in this regard : a similar thing happened in the Industrial Revolution, and again in the 'information revolution' with the advent of computers and the internet.

In the end, artificial intelligence should not be feared. AI is being created to help us, and like all future technologies, we need to ensure that it is used appropriately.

- Question 8 : what does “downsides” mean ?
- Question 9 : Can you explain this sentence : “If we train AI with biased data then the AI will also be biased”.  
Give an example of such an issue (I mean by using biased data).
- Question 10 : What does this article mean by “we need to ensure that it is used appropriately” ?

## VI - TP

### Exercice 6 Traitement et représentation d'une série de données

On souhaite étudier l'évolution des prix du pétrole depuis 1990.

1. Dans l'espace de travail EcoleDirecte, télécharger les données sous forme d'un fichier .csv : SNT\_prix\_petrole\_dollar.csv
2. Importer le fichier csv dans votre logiciel tableur préféré.
3. Repérer les descripteurs de ce fichier.
4. Représenter graphiquement l'évolution des prix du baril de pétrole depuis 1990.
5. À quelle date le prix du pétrole a-t-il été le plus élevé ?
6. Calculer le prix moyen du baril de pétrole sur toute la période considérée.

### Exercice 7 Traitement des données

On souhaite étudier l'évolution des prix du pétrole depuis 1990.

1. Dans l'espace de travail EcoleDirecte, télécharger les données sous forme d'un fichier .csv : SNT\_adherents\_association.csv
2. Identifier le format de ce fichier.
3. Importer le fichier csv dans votre logiciel tableur préféré.
4. Combien d'objets compte ce fichier ? Quels sont ses descripteurs ?
5. Préciser selon quel descripteur les adhérents sont triés dans le fichier.
6. Déterminer la moyenne d'âge des adhérents de cette association.
7. À l'aide d'une opération de tri, déterminer le nom du membre apparaissant en 3e position dans l'ordre alphabétique des prénoms.

### Exercice 8 TP Python

Paul possède une collection de fichiers audios au format MP3. Il a exactement 62452 fichiers, et il sait qu'en moyenne un fichier MP3 à une taille de 4,5 Mo. Il souhaiterait savoir :

- Combien de carte perforées sont nécessaires pour stocker toute sa musique ?
- Combien de disquette 3 pouces 1/2 sont nécessaires pour stocker toute sa musique ?
- Combien de CDROM sont nécessaires pour stocker toute sa musique ?
- Combien de clé USB 16 Go sont nécessaires pour stocker toute sa musique ?
- Combien de clé disque dur 250 Go sont nécessaires pour stocker toute sa musique ?
- Combien de data center de 300000 m<sup>2</sup> sont nécessaires pour stocker toute sa musique ?

Voici quelques éléments concernant les supports de stockage de données qui vous seront utiles pour la suite de ce TP.

<b>Unité de stockage</b>	<b>Capacité de stockage</b>
Carte perforée	960 octets
Disquette 3 pouces 1/2	1,44 Mo
CD-ROM	650 Mo
Clé USB de 16 Go	16 Go
Disque Dur de 250 Go	250 Go
Data center (300000 m <sup>2</sup> )	1 Yo

1 octet (o)	8 bits
1 Téraoctet ( To )	1000 Go = $10^{12}$ o
1 Pétaoctet ( Po )	1000 To = $10^{15}$ o
1 Exaoctet ( Eo )	1000 Po = $10^{18}$ o
1 Zetaoctet ( Zo )	1000 Eo = $10^{21}$ o
1 Yotaoctet ( Yo )	1000 Zo = $10^{24}$ o

## PARTIE 1 : Calculs

1. Calculez le volume total, en octets, de la collection de fichiers audios de Paul.
2. Pour chacun des supports de stockage de données (tableau de gauche), convertir leur capacité de stockage en octets. Vous exprimerez les résultats en écriture scientifique.
3. Pour chacun des supports de stockage, calculer le nombre de supports nécessaire pour stocker la collection de fichiers audios de Paul.

## PARTIE 2 : Programmation Python

1. Dans l'éditeur python de Basthon, copiez le code suivant et interpréter la première ligne de ce code Python.

```
a = 2.5e5
print(a)
```

2. Dans votre éditeur Basthon, copiez le code suivant.

```
def nb_unite_stockage(volume_donnee, capacite_unite_stockage):
    nb_unite_stockage = ...
    return int(nb_unite_stockage)

volume_total = ...
capacite_carte = 960
capacite_disquette = 1.44e6
capacite_cdrom = ...
#a completer
print( nb_unite_stockage(volume_total,capacite_carte) )
#a completer
```

3. Complétez le code Python et affectez le volume total de la collection de fichiers audios à la variable `volume_total`.
4. Pour chacunes des valeurs calculées à la question 2 de la partie 1, créez une variable et affectez la valeur correspondante en octets en utilisant l'écriture scientifique. Le code python propose déjà deux variables `capacite_carte` et `capacite_disquette`.
5. Le code python implémente une fonction que vous devez compléter.
  - Quelle est le nom de la fonction ?
  - Quels sont les paramètres de cette fonction ?
  - Complétez le code de la fonction afin qu'elle retourne le nombre d'unité de stockage nécessaire pour stocker une quantité de donnée dont le volume est donné par le paramètre `volume_donnee` et la capacité de stockage de l'unité est donnée par le paramètre `capacite_unite_stockage`.

6. L'avant dernière ligne de ce code Python permet d'appeler la fonction `nb_unite_stockage` pour afficher le nombre de carte perforée nécessaire pour stocker la collection de fichiers audios. Vérifiez que vous obtenez le même résultat que lors de vos calculs de la question 3, partie 1.
7. Complétez le code Python pour afficher le nombre d'unité de stockage pour chacun des type de support. Comparez avec vos résultats de la question 3, partie 1.

