Les données structurées et leur traitement

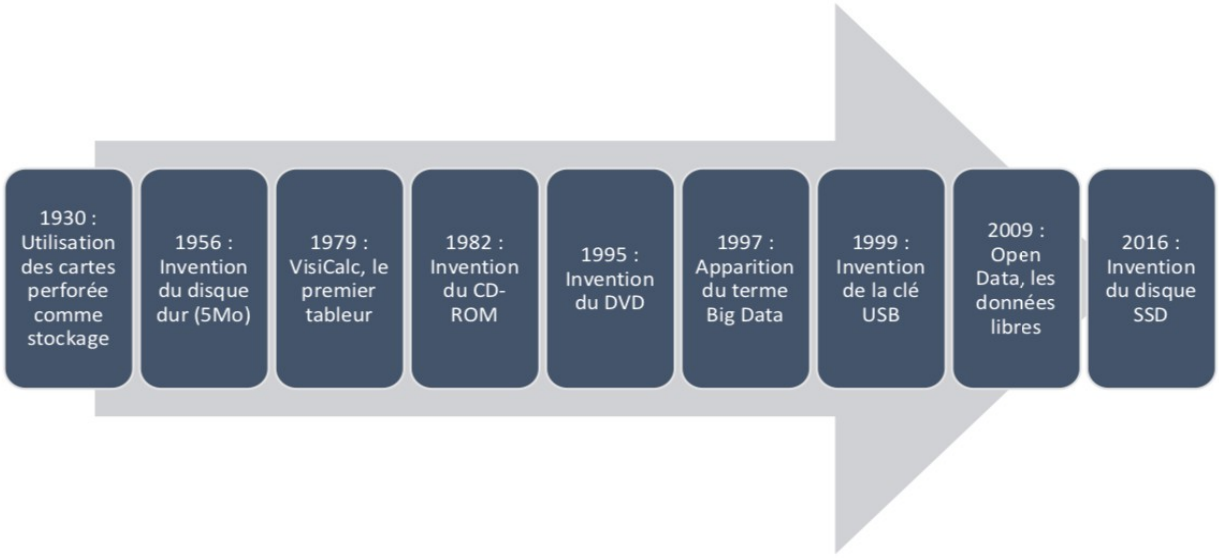
Mots clés du chapitre	Le programme
Données structurées Descripteur d'un objet Données ouvertes Traitement des données Métadonnées Le cloud Le big data Les Data Centers	Identifier les principaux formats de données Identifier les différents descripteurs d'un objet Réaliser des opérations de recherche, filtre ou calcul sur une ou plusieurs tables Retrouver les métadonnées d'un fichier personnel Connaître les problèmes induits par le traitement et stockage des données

« Chaque minute en 2019, on a recensé dans le monde :

100 000 tweets
3 800 000 recherche Google
190 000 000 mails envoyés
30 h de vidéo mis sur YouTube
1 000 000 de connexions facebook

»

I- Un peu d'histoire



II- Quelques définitions

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 .

Plusieurs descripteurs peuvent alors être utiles pour décrire un même objet (par exemple, des descripteurs permettant de caractériser un contact : nom , prénom, numéro de téléphone)

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 →

Une valeur du descripteur « Nom » →

Un objet →

Nom	Capitale	Hymne	Superficie (km²)
France	Paris	La Marseillaise	632 734
Chine	Pékin	La Marche des volontaires	9 596 961
USA	Washington	The Star-Spangled Banner	9 833 517
Argentine	Buenos Aires	Himno Nacionnal Argentino	2 791 810

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

Si il est évident qu'une donnée est personnelle quand le nom et/ou prénom est indiqué, il en est de même si on donne une adresse IP, une photographie, un numéro de téléphone, une donnée biométrique, des données de localisation (comme latitude et longitude) etc … . Le RGPD (voir ci dessous) est alors là pour encadrer ces données .

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.

A tout fichier 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 puis *lire les informations*

III- Les formats de fichiers de données

Pour assurer la persistance des données, ces dernières sont stockés dans des fichiers. Dans l'open data, les deux formats les plus utilisés sont le format CSV et le format JSON

1- Le 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 entre les valeurs suivantes dans 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 qui donne après l'avoir enregistré au format .csv le tableau suivant :

	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					

2- Le 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 très utilisé sur internet : le JavaScript.

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

Exemple :

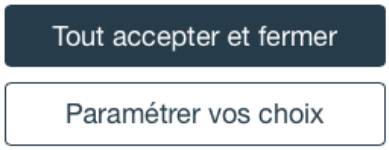
On entre les valeurs suivantes dans un traitement de texte

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

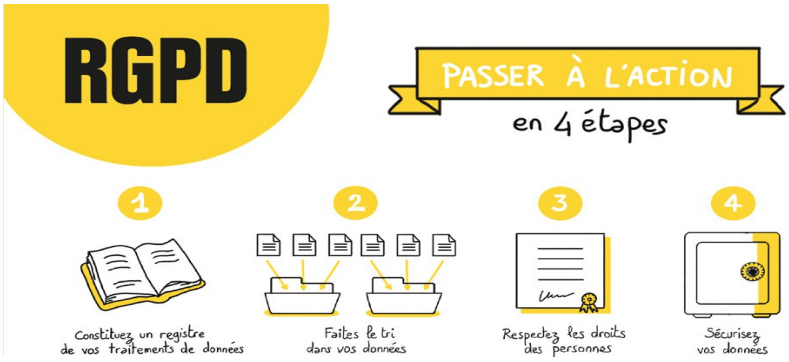
Ce qui donne après l'avoir enregistré au format .json :

IV- Impact sur les pratiques humaines

1- Les données personnelles et le RGPD



Même si certaines de ces données sont dites ouvertes, leur producteur considérant qu'il s'agit d'un bien commun , on assiste aussi au développement d'un marché de la donnée où des entreprises collectent et revendent des données sans transparence pour les usagers. Il a donc fallu créer un cadre juridique permettant de protéger les usagers, préoccupation à laquelle répond le Règlement Général sur la Protection des Données : le RGPD



2- Le BIG DATA

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

3- Le cloud

Le cloud désigne l'ensemble des ressources informatiques (stockage , service) 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. Ainsi, les ressources sont dites « dans le nuage » qui représente le vaste réseau internet.

4- Impact 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. D'une manière générale, l'émergence du cloud est liée à l'utilisation de data centers.

De telles installations dégagent cependant énormément de chaleur et doivent être refroidies pour éviter toute panne, ce qui induit une consommation électrique très élevée. **On estime qu'environ 20 % de la consommation électrique liée au numérique provient des data centers.**

Annexe

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

A noter : C'est depuis 1998 que l'IEC (une commission internationale) a décidé de prendre cette norme pour tous (1 Ko = 1000 o) . Cependant, au commencement, pour des raisons binaires, on avait 1 Ko = 1024 o . Pour ne pas bouleverser les usages, la commission a introduit de nouveaux préfixes binaires : le kibi (noté Ki), le mébi (noté Mi), le gibi (noté Gi), etc. permettant de retrouver les puissances de 2 traditionnelles...

Officiellement on a donc : 1 Kio = 1 024 o ; 1 Mio = 1 024 Kio = 1 048 576 o ; 1 Gio = 1 024 Mio = 1 048 576 Kio = 1 073 741 824 o...