Enseignement SNT

LYCÉE de l'ÉLORN
le choix de votre réussite

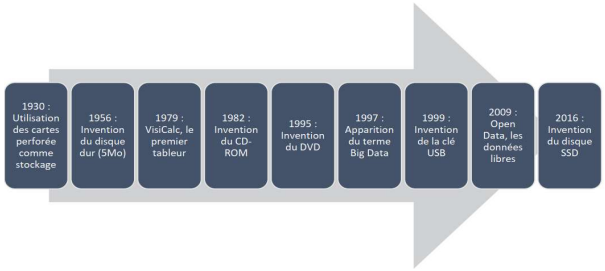
Les données structurées

Reprise du support de cours rédigé par Jean-Christophe Bonnefoy - http://www.jcbmathsandco.fr/ -

1. Repères historiques, volumes de données

Chaque minute en 2020, il s'est échangé dans le monde :

- *100 000 tweets*
- *3 800 000 de Google*
- *190 000 000 de mails envoyés*
- *30h de vidéos mis sur Youtube*
- *1 000 000 de connexions Facebook*
- *1 000 000 000 de dollars d'achats en ligne*
- *…*



2. Une donnée, c'est quoi ?

Une ***donnée*** *(*data *en anglais) est*

Une donnée peut être élémentaire ou complexe.

- Une donnée *élémentaire* représente une caractéristique de base (un nom, un numéro, etc.). Cette donnée est caractérisée par un *descripteur* qui permet de donner le format dans lequel cette donnée est représentée.
- Une donnée *complexe* est constituée de plusieurs données élémentaires.

Prenons le cas d'une adresse postale. Il s'agit d'une donnée complexe constituée de 8 données élémentaires. Lister ces données et caractériser leur descripteur.

M. Archibald TARTEMPION

43 Rue des Jolis Arbres

29280 LANDERNEAU

3. Les données personnelles et leurs protections

Une ***donnée personnelle*** est

Les données personnelles sont protégées dans tous les états membres de l'Union Européenne par une *loi Informatique et libertés*. En effet, depuis 2018, ... oblige tout organisme qui collecte des données à prouver la nécessité de cette collecte, à protéger les données recueillies et à être plus transparent sur leurs utilisations. En France, l'autorité compétente est la ..... Elle est chargée de veiller à la protection de l'identité humaine, des droits de l'homme, de la vie privée et des libertés individuelles.

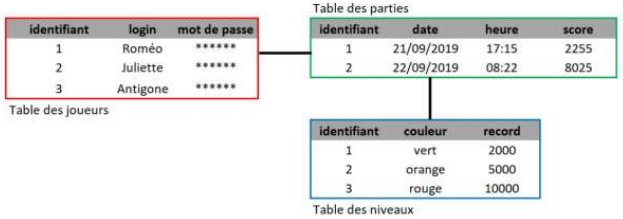
4. La structuration des données

Les données utilisées pour une application sont souvent très nombreuses. Il est donc nécessaire de les organiser pour pouvoir les utiliser de manière efficace. Lorsque des données partagent les mêmes descripteurs, on les rassemble dans une ***collection***.

On les présente souvent sous forme de tables dont les colonnes représentent les descripteurs et les lignes représentent les données. La valeur de cette donnée se lit ainsi à l'intersection de la ligne et de la colonne.

- Lorsque l'on peut classer les données sous la forme de tables, on parle de *données structurées*.
- Lorsque l'on regroupe des collections de données reliées entre elles, on parle de *base de données*. Une base de données est représentée dans un format spécifique qui fait apparaitre les collections (sous forme de tables) et les liens entre ces collections (appelées relations)

Série	Titre de l'album	Dessinateur	Editeur	Année de parution
Tintin	Tintin au Tibet	Hergé	Casterman	1959
Astérix le gaulois	La serpe d'or	Uderzo	Dargaud	1962
Lefranc	La grande menace	Martin	Le Lombard	1954
Blake et Mortimer	La marque jaune	Jacobs	Le Lombard	1956



5. Les formats de fichiers des données structurées

Pour assurer la persistance des données, ces dernières sont stockées dans des fichiers. Il en existe une multitude, cependant les deux formats les plus utilisés sont le format **CSV** et le format **JSON**.

5.1. Le format CSV (Comma-Separated-Value)

Dans un fichier

Les caractères les plus connus sont la virgule, le point-virgule (comme le montre la figure ci-contre) ou encore la tabulation.

```
Nom ;prénom ;login ;mot de passe ;service
MARTIN ;Louis ;ML ;******* ;Administratif
DURAND ;Claire ;DC ;******* ;Commercial
DUPOND ;Georges ;DG ;******* ;Technique
DUBOIS ;Odile ;DO ;******* ;Direction
```

	A	B	C	D	E
1	Nom	prénom	login	mot de passe	service
2	MARTIN	Louis	ML	******	Administratif
3	DURAND	Claire	DC	******	Commercial
4	DUPOND	Georges	DG	******	Technique
5	DUBOIS	Odile	DO	******	Direction
6					
7					

Le même fichier CSV mis sous forme de tableau

5.2. Le format JSON (JavaScript Objet Notation)

Dans un fichier

L'intérêt du format JSON est qu'il permet un stockage de données plus complexes que celles présentes dans un fichier CSV. Il associe des paires de descripteur/valeur séparées par le caractère « : », et chaque paire est séparée de la suivante par le caractère « , ».

```
{
"Espèce" : "Chien",
"Age" : 6,
"Race" : "Golden Retriever",
"Trait" : {
            "CouleurYeux" : "Marron",
            "CouleurPelage" : "Jaune"
          }
}
```

Remarque : la quatrième valeur de l'exemple ci-dessus est une donnée imbriquée qui comprend 2 paires de descripteur/valeur. On peut ainsi imbriquer les objets à l'infini contrairement aux formats précédents.

6. Les supports de stockage

Les fichiers de données sont stockés physiquement sur un support. Il en existe une multitude.

- Le disque dur : stockage mécanique d'environ 1 To aujourd'hui.
- Les disques CD et DVD : supports optiques de 700 Mo pour les CD et jusqu'à 17 Go pour les DVD multicouches.
- La clé USB : support externe de type flash (rapidité d'accès aux données) qui possède un bon rapport prix / capacité mais peu fiable
- Les cartes mémoires (SD, microSD,…) : même technologie que les clés USB mais surtout adapté au monde de la photographie et aux smartphones.
- Le disque SSD : stockage de capacité équivalente au disque dur mécanique et de grande rapidité (technologie flash), c'est le successeur des « vieux » disques durs.

7. Le traitement des données structurées

Le *traitement des données* consiste à

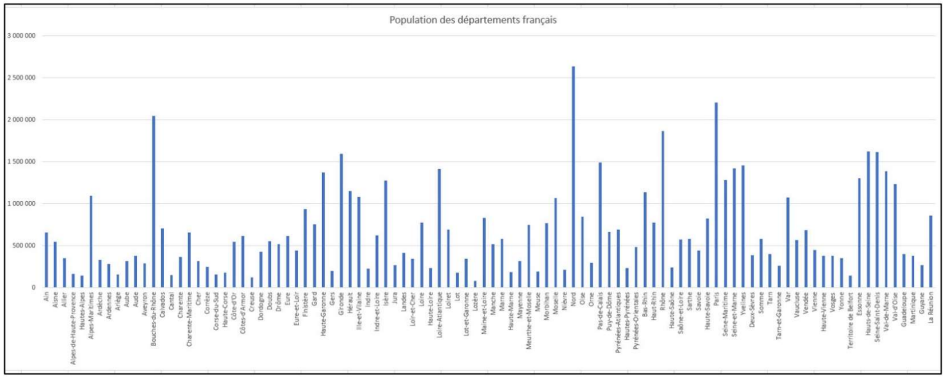
Ces actions peuvent être de nature complexe, il s'agit cependant dans la majorité des cas d'actions élémentaires comme l'affichage, la recherche, le tri et le filtrage.
Dans cette partie, on prendra comme données, le fichier du recensement de la population française en vigueur au 1er janvier 2019 issu de l'Institut National de la Statistique et des Etudes Economiques (INSEE).

	A	B	C	D	E	F	G	H	I
1	Code région	Nom de la région	Code département	Nom du département	Nombre d'arrondissements	Nombre de cantons	Nombre de communes	Population totale	Superficie (en km²)
2	84	Auvergne-Rhône-Alpes	01	Ain	4	23	407	655 171	5 762
3	32	Hauts-de-France	02	Aisne	5	21	804	549 587	7 369
4	84	Auvergne-Rhône-Alpes	03	Allier	3	19	317	349 336	7 340
5	93	Provence-Alpes-Côte d'Azur	04	Alpes-de-Haute-Provence	4	15	198	167 331	6 925
6	93	Provence-Alpes-Côte d'Azur	05	Hautes-Alpes	2	15	163	146 148	5 549
7	93	Provence-Alpes-Côte d'Azur	06	Alpes-Maritimes	2	27	163	1 098 539	4 299
8	84	Auvergne-Rhône-Alpes	07	Ardèche	3	17	339	334 591	5 529
9	44	Grand Est	08	Ardennes	4	19	452	283 004	5 229
10	76	Occitanie	09	Ariège	3	13	331	158 205	4 890
11	44	Grand Est	10	Aube	3	17	431	316 639	6 004
12	76	Occitanie	11	Aude	3	19	436	377 580	6 139
13	76	Occitanie	12	Aveyron	3	23	285	289 481	8 735
14	93	Provence-Alpes-Côte d'Azur	13	Bouches-du-Rhône	4	29	119	2 047 433	5 087
15	28	Normandie	14	Calvados	4	25	537	709 715	5 548
16	84	Auvergne-Rhône-Alpes	15	Cantal	3	15	247	151 615	5 726
17	75	Nouvelle-Aquitaine	16	Charente	3	19	381	365 697	5 956
18	75	Nouvelle-Aquitaine	17	Charente-Maritime	5	27	466	660 458	6 864
19	24	Centre-Val de Loire	18	Cher	3	19	290	315 100	7 235
20	75	Nouvelle-Aquitaine	19	Corrèze	3	19	283	249 707	5 857
21	94	Corse	2A	Corse-du-Sud	2	11	124	156 958	4 014
22	94	Corse	2B	Haute-Corse	3	15	236	179 037	4 666
23	27	Bourgogne-Franche-Comté	21	Côte-d'Or	3	23	704	546 466	8 763
24	53	Bretagne	22	Côtes-d'Armor	4	27	355	618 478	6 878
25	75	Nouvelle-Aquitaine	23	Creuse	2	15	258	123 500	5 565

Extrait du fichier contenant ces données

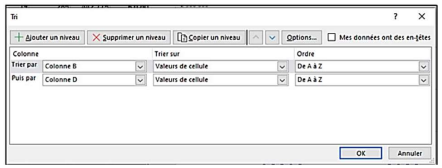
7.1. L'affichage adapté

Problématique : Quel est le département le plus peuplé ?
De façon classique, on visualise les données par les graphiques : histogrammes et diagrammes circulaires. On utilise pour cela le tableur avec l'outil « insertion de graphiques ».



7.2. Le tri des données

Le **tri** dans une table consiste à modifier l'ordre des données pour qu'elles soient présentées dans un ordre croissant ou décroissant selon le choix d'un ou plusieurs critères.



7.3. Le filtrage des données

Le **filtrage** dans une table consiste à sélectionner des données contenant une information particulière afin de n'afficher que ces données-là.

7.4. Réaliser des calculs à partir des données

A partir des données existantes, on peut calculer d'autres données. Cela peut être générer avec les fonctions disponibles sur le tableur (en VisualBasic ou en Python) ou en important les fichiers sur un Script Python.

```
fichier = open("population_superficie_departements.csv", "r")
ligne_nom_colonne = fichier.readline()
poptotal = 0
ligne_donnee = fichier.readline()
lst = []
while ligne_donnee != "":
    lst = ligne_donnee.split(";")
    poptotal = poptotal + int(lst[7])
    ligne_donnee = fichier.readline()

fichier.close()
print("La population totale de la France est de ",poptotal, "habitants")
```

8. Le Big Data

Le ***big data*** désigne

L'expression « Big Data » date de 1997 et répond à la ***règle des 5V*** :

- Le ***Volume*** de données considérable à traiter
- Une grande ***Variété*** d'informations (venant de diverses sources, non-structurées, organisées, Open, etc.)
- Un certain niveau de ***Vélocité*** à atteindre (les données sont produites, récoltées et analysées en temps réel).
- La ***Véracité*** concerne la fiabilité et la crédibilité des informations collectées.
- La ***Valeur*** correspond au profit qu'on peut tirer de l'usage du Big Data

9. Le Cloud

Le ***cloud*** provient du terme

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.

10. Les Data Centers et l'impact environnemental

Un ***data center*** est un

L'intérêt majeur des data centers est de garantir la haute disponibilité d'une grande quantité de données dans des conditions optimales de sécurité.

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

Si Internet était un pays, il serait le 3ème plus gros consommateur d'électricité au monde derrière la Chine et les Etats-Unis.

10. Application

On a enregistré les données d'un répertoire téléphonique au format CSV et JSON.

```
"nom";"numéro";"adresse"
'Pierre';'0614122178';'5, rue du fort'
'Pauline';'0612166114';'18, rue de la poste'
'Kilian';'0664122432';'2, avenue des oeillets'
```

```
{
        {
                'nom':'Pierre',
                'numéro':'0614122178',
                'adresse':'5, rue du fort'
        },
        {
                'nom':'Pauline',
                'numéro':'0612166114',
                'adresse':'18, rue de la poste'
        },
        {
                'nom':'Kilian',
                'numéro':'0664122432',
                'adresse':'2, avenue des oeillets'
        }
}
```

1/ Quels sont les descripteurs du répertoire téléphonique ?

2/ Quelles sont les différentes valeurs du descripteur nom ?

3/ Nous avons récupéré les données suivantes, présentées sous forme de tableau, sur des aliments.
Ecrire les fichiers CSV et JSON correspondants.

Aliment	Nutriscore	Calories/100g
gaufre	C	291
Yaourt	A	65
Muesli	A	175
Pain de mie	B	234