Fiche

I. Données

• Les données sont essentielles pour toute activité numérique (gestion de stock, fiche client...).

• Une **donnée** est un élément (chaîne de caractères, nombre, image...) décrivant un objet (individu, société, événement, machine, fichier...). Par exemple, la date de création de la fondation Abbé-Pierre est 1992, c'est une donnée. Une donnée peut-être créée par un individu ou par un appareil.

• Une **donnée personnelle** est une information se rapportant à une personne physique identifiée ou identifiable par référence à des éléments qui lui sont propres (nom, numéro de sécurité sociale...).

II. Données structurées

• Structurer correctement des données permet de les utiliser et exploiter aisément afin de produire de l'information. Il faut donc organiser et classer les données.

• Plusieurs **descripteurs** peuvent être utiles pour décrire un même objet. Un passeport français contient plusieurs descripteurs comme le numéro du passeport, le nom, les prénoms, le sexe, la taille, la couleur des yeux, la date de naissance, le lieu de naissance...

• Par exemple, sur le passeport de l'acteur français Pierre Richard :
- à côté du **descripteur** lieu de naissance, il est inscrit la **valeur** Valenciennes ;
- à côté du **descripteur** date de naissance, il est inscrit la **valeur** 16/08/1934.

• Une **collection de données** est un moyen de regrouper de manière **structurée** des objets partageant les mêmes descripteurs. Elle est généralement représentée sous la forme d'une **table** : les descripteurs en colonne, les objets en ligne et les valeurs dans les cellules situées à l'intersection.

• Voici une collection de données contenant deux objets. Quatre descripteurs sont utilisés et huit valeurs sont visibles.

Nom	Prénom	Date de naissance	Lieu de naissance
Macron	Emmanuel	21 12 1977	Amiens
Hollande	François	12 08 1954	Rouen

• Les tableaux peuvent être stockés dans un fichier avec un format spécifique comme le format **csv** (*Comma-separated values*). Dans un tel fichier, les données sont sous un format texte simple et séparées les unes des autres par un caractère (la virgule par exemple). Chaque ligne correspond à une ligne du tableau.

• On peut ouvrir un fichier csv avec un éditeur de texte basique. Voici un exemple :
Nom,Prénom,Date de naissance,Lieu de naissance
Macron,Emmanuel,21 12 1977,Amiens
Hollande,François,12 08 1954,Rouen

• Le caractère de séparation choisi ne doit pas figurer dans les valeurs sinon un logiciel ne pourra pas identifier correctement les descripteurs.

• Le site web data.education.gouv.fr propose des **données ouvertes** brutes et en libre accès (*OpenData*) et permet de télécharger des collections de données au format csv. Des filtres sont disponibles sur le côté gauche pour pouvoir affiner la collection spécifique recherchée.

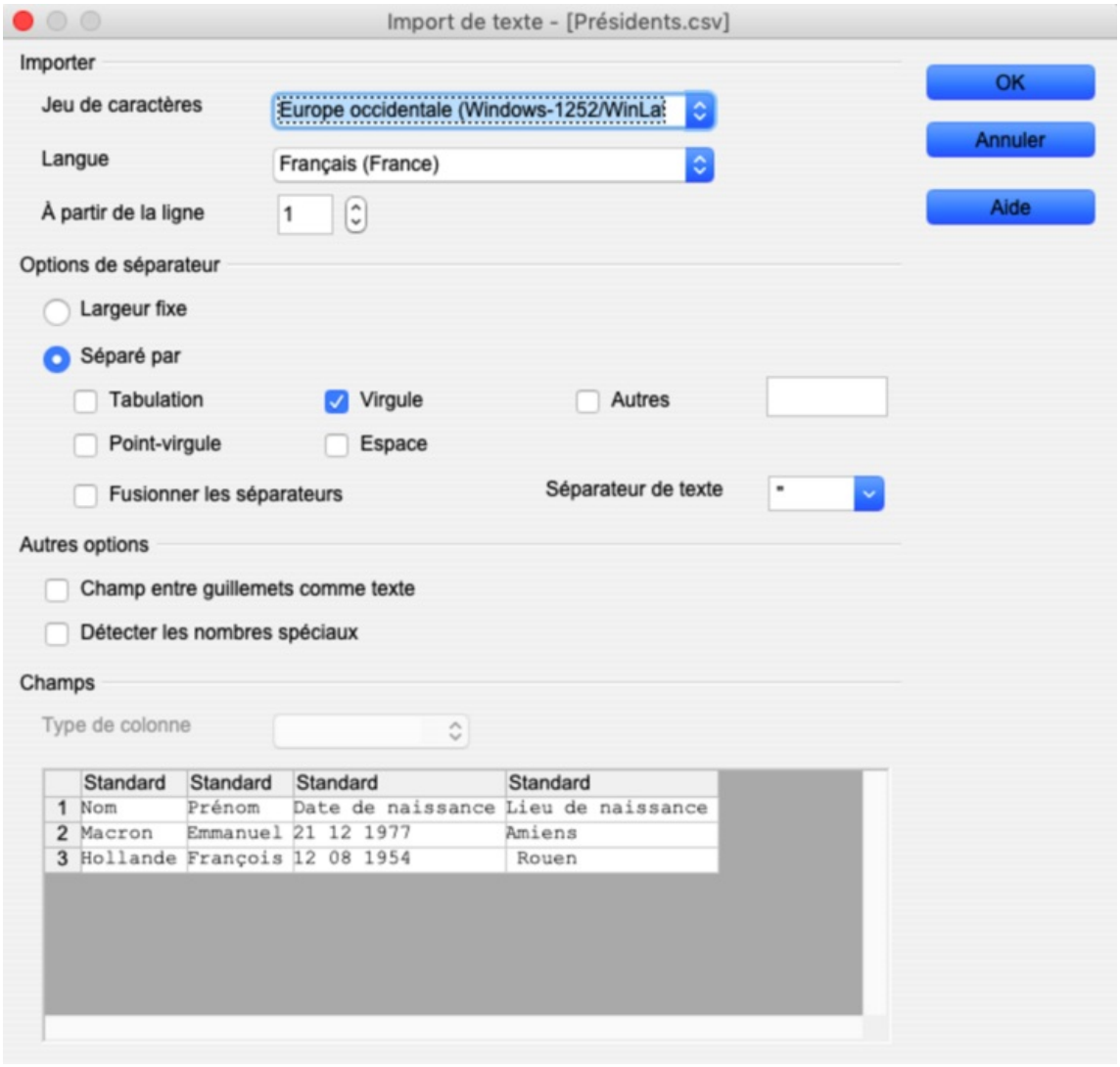
• Les sites web data.sncf.com, data.gouv.fr, data.bnf.fr proposent de la même manière des collections de données dans certains formats.

III. Traitement de données structurées

Tableur

• On peut ouvrir un fichier csv avec un tableur qui devrait (après lui avoir indiqué que le séparateur utilisé est la virgule par exemple) utiliser une colonne par descripteur et une ligne par objet.

• Dans le cas du logiciel gratuit OpenOffice (openoffice.org), après avoir ouvert une feuille de calcul il faut sélectionner « Insertion » puis « Feuille à partir d'un fichier ». On sélectionne alors le fichier csv à importer.



• Selon le pays, la langue et les caractères utilisés on choisira UTF-8 ou un autre (on peut prévisualiser dans le bas de la fenêtre pour éviter qu'un « é » soit supprimé ou transformé en « © »).

• Ici le séparateur « , » a été repéré directement par le logiciel, mais il faut évidemment vérifier que cela est correct.

• Après validation, on choisit l'endroit de la feuille de calcul pour affichage.

• Une fois importées, on peut **trier les données** : on va modifier l'ordre des objets.

• Il suffit de sélectionner « Données » puis « Trier... » et de choisir un descripteur. On peut choisir un tri dans l'ordre croissant ou décroissant. On peut même choisir d'autres descripteurs pour effectuer un tri dans un tri ! Pour un tri simple, on peut aussi directement sélectionner la colonne voulue et choisir le bouton « A->Z » pour l'effectuer dans l'ordre croissant.

• On peut aussi **filtrer** les données, c'est-à-dire afficher les objets qui contiennent une valeur particulière. Il suffit de sélectionner « Données » puis « Filtre ».

• On peut alors choisir « AutoFiltre » qui va mettre en place automatiquement autant de filtres que de valeurs différentes par descripteur.

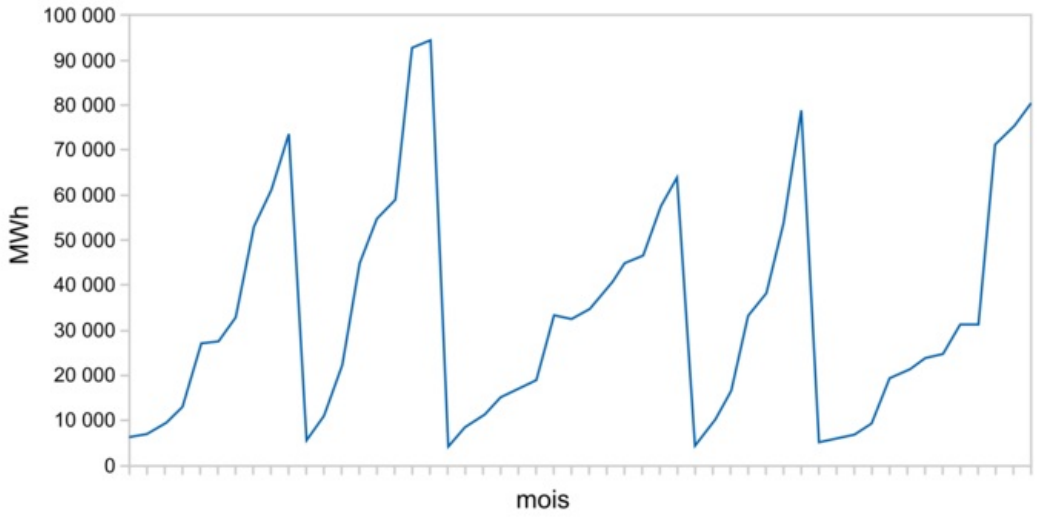
• On peut choisir « Filtre standard » afin de sélectionner soi-même un descripteur particulier et de n'afficher que certaines valeurs (conditions à choisir).

• Enfin on peut mettre en place des filtres sur toutes les colonnes ou uniquement sur certaines.

• Les outils « tri » et « filtres » sont particulièrement utiles lorsqu'une collection de données contient des centaines ou des milliers de lignes.

• On peut représenter certaines données graphiquement. Il suffit de sélectionner « Insertion » puis « Diagramme ». Ensuite :
1. Choisir le type de diagramme souhaité : diagramme en barres, nuages de points...
2a. Sélectionner la plage des données. En général, il s'agit de toute la feuille de calcul, donc on écrit $Feuille1_2.$A$1:$F$53 si la dernière cellule non vide est la cellule F53.
2b. Choisir « Série de données en colonnes ».
2c. Cocher « Première ligne comme étiquette ».
3. Choisir la (ou les) série(s) de données à représenter. On supprime en fait les colonnes non voulues.
4. Légender le diagramme, les axes...

• On peut ensuite en double cliquant sur les axes, les points, jouer sur le côté esthétique (couleur, épaisseur...). Voici un exemple à partir du fichier production-mensuelle-biomethane.csv obtenu sur data.gouv.fr.



• La visualisation de données (dataviz) est un bon outil de communication, mais elle peut parfois induire les individus en erreur.

Python

• La bibliothèque *pandas* permet d'importer des fichiers csv et de manipuler des collections de données. Considérons le fichier production-mensuelle-biomethane.csv obtenu sur le site web data.gouv.fr. Les trois premières lignes de ce fichier sont :
Année;Mois;Nombre de sites biométhane;Production mensuelle de biométhane (MWh);Nombre de sites biométhane GRTgaz;Production mensuelle sur réseau GRTgaz (MWh)
2015;2015-07;11;6334;;
2015;2015-08;12;6982;;

• Il y a six descripteurs, mais les noms de certains sont un peu longs : on pourra donc les renommer. À partir de la collection initiale (variable nommée « table » dans ce programme), on peut créer d'autres collections.

```
import pandas
from pandas import DataFrame, read_csv
import matplotlib
liste=["Année","Mois","Sites","Prod mensuelle","Sites GRT","Prod mensuelle GRT"]
table=read_csv("production-mensuelle-biomethane.csv",sep=";",names=liste,skiprows=1)
print(table.to_string())
extrait=table[liste[3]]
extraitGRT=table[table[liste[4]] >0]
print(extraitGRT.to_string())
sup50=table[table[liste[2]]+table[liste[4]]==45]
print(sup50.to_string())
ordrecroissant=table.sort_values("Prod mensuelle")
print(ordrecroissant.to_string())
max=table.max()
print(max.to_string())
troispremlignes=table.head(3)
print(troispremlignes.to_string())
graphique=extrait.plot(kind="line")
matplotlib.pyplot.savefig("Evolution.pdf")
```

• Par exemple :

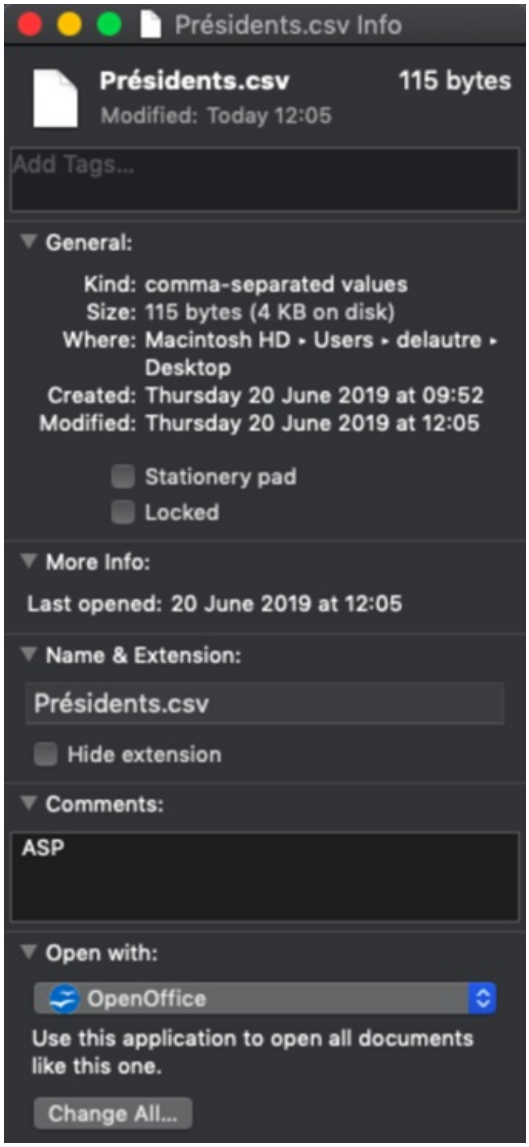
- la variable « extraitGRT » est la collection des données pour lesquelles le nombre de sites de GRTgaz est strictement positif ;
- la variable « sup50 » est la collection des données pour lesquelles le nombre total de sites est égal à 45 ;
- la variable « ordrecroissant » est la collection des données rangées dans l'ordre croissant selon les productions mensuelles.

• On peut également construire une représentation graphique représentant l'évolution de la production mensuelle de biométhane à l'aide de la bibliothèque matplotlib et créer un fichier PDF.

IV. Métadonnées

• À tout fichier sont associées des **métadonnées** qui permettent d'en décrire le contenu (date de création, taille, auteur...). Elles peuvent être complétées automatiquement (par l'appareil) ou renseignées manuellement par un individu.

• Pour les observer, sur Windows, il suffit de sélectionner un fichier et à l'aide d'un clic droit on peut sélectionner « Propriétés » (ou « Get Info » sur macOS).



• Le fichier Présidents.csv a été créé le 20 juin 2019 et sa taille est de 116 octets.

• Certaines métadonnées d'un fichier peuvent parfois être modifiées ou supprimées. Pour cela, sur Windows, cliquer sur l'onglet « Détails », puis en bas sur « Supprimer les propriétés et les informations personnelles ».

• Les métadonnées sont souvent contenues dans le fichier lui-même.

• Un fichier audio contient souvent des métadonnées (utilisant le standard ID3v2) comme artistes, album, genre, compositeur, paroles...

V. Données dans le nuage (*cloud*)

• Afin de permettre la réutilisation de données et de les protéger d'éventuels dommages, il est nécessaire de bien les conserver. On peut utiliser un stockage interne (HDD ou SSD) ou bien externe : clef USB, disque externe, CD, DVD ou encore le *cloud*.

• Le ***cloud*** est un espace informatique de stockage de données sur des serveurs connectés à Internet ; ce service gratuit ou payant est proposé par diverses entreprises. On peut ainsi accéder à ses données depuis n'importe quel endroit et de n'importe quel appareil tant que l'on dispose d'une connexion à Internet.

• On peut également **synchroniser** certains dossiers de son appareil avec le *cloud* afin de ne pas avoir à le faire soi-même de manière systématique après des modifications dans un document, un ajout de fichier dans un dossier ou une réorganisation des fichiers. Toute modification dans le dossier de son appareil sera alors appliquée à sa copie dans le *cloud*.

• On peut autoriser certains utilisateurs à avoir accès (lecture et/ou écriture) à un fichier ou un dossier sur le *cloud*, ce qui peut être utile pour un travail collaboratif.

• Les centres de données (***data centers***) sont les lieux principaux de stockage des données du *cloud*. Ils comportent des ordinateurs, des serveurs et des baies de stockage qui sont connectés à un réseau (Internet par exemple) et fonctionnent en permanence.

• Ils sont **très gourmands en électricité**. Celle-ci est utilisée pour l'alimentation, mais aussi le refroidissement des serveurs. En 2015 les centres de données du monde entier ont consommé 416 TWH d'énergie soit presque autant que la France. En 2017, l'écosystème numérique représentait environ **7 % de la consommation mondiale d'électricité**. Cependant, ils permettent également des économies d'échelle grâce à la mutualisation des ressources (maintenance, matériel...) entre les utilisateurs. En 2019, il y a plus de 400 *hyperscale data centers* dans le monde. Pour restreindre cette consommation, chacun peut éteindre complétement ses appareils la nuit et supprimer des courriels inutiles.

• Enfin des métaux rares (ressources limitées dont l'extraction est polluante) sont utilisés dans la fabrication des serveurs.

VI Base de données

• Afin d'éviter des répétitions, on peut stocker certaines données dans une collection à part. **Une base de données** regroupe plusieurs collections de données reliées entre elles. Par exemple, les annuaires pages jaunes des départements français forment une base de données.

• Une recherche dans une base de données peut ainsi croiser des collections différentes sur un même descripteur.

Big Data

De nos jours, il y a une surabondance de données. Le ***Big Data*** désigne à la fois ces données massives et le domaine qui travaille sur le traitement de ces données dans le but d'**extraire de l'information pertinente** (corrélation). Des applications sont utilisées par certains athlètes ou équipes pour analyser leurs adversaires ou leur propre pratique. Les applications politiques ne sont pas encore au point (données de qualité moyenne, algorithme désuet). Par exemple, les sondages des élections américaines du 8 novembre 2016 étaient fort éloignés de la réalité du résultat. Des applications dans le domaine de la santé existent comme la recherche de facteurs environnementaux ou génétiques dans l'apparition d'une maladie.

Surveillance

En Chine, dans la région du Guangdong et sur l'ensemble du territoire d'ici la fin 2020, le crédit social (*social credit system*) vise à évaluer la vertu et la bonne foi des individus et à les classer en bons ou mauvais citoyens. La note d'un citoyen est calculée dans ce système à partir des données dont l'État dispose sur lui.

Transparence

• Des entreprises (Facebook par exemple) collectent des données sur leurs utilisateurs et les utilisent de manière plus ou moins transparente. C'est à l'utilisateur de lire avec attention ce que l'entreprise collecte et ce qu'elle va faire de ces informations. Il est par exemple écrit clairement dans la **politique d'utilisation des données** de Facebook :

• « Nous recueillons le contenu, les communications ainsi que d'autres informations que vous fournissez lorsque vous utilisez nos Produits. Cela peut comprendre des informations présentes dans le contenu que vous fournissez (par exemple, des métadonnées) ou concernant un tel contenu, telles que le lieu d'une photo ou la date à laquelle un fichier a été créé. Si vous utilisez nos Produits pour effectuer des achats ou toute autre transaction financière (par exemple, lorsque vous effectuez un achat dans un jeu ou lorsque vous faites un don), nous recueillons des données concernant cet achat ou cette transaction. Ceci comprend vos informations de paiement, telles que le numéro de votre carte de crédit ou de débit et d'autres informations concernant votre carte, d'autres informations de compte et d'authentification, ainsi que des données de facturation et de livraison, et vos coordonnées. »

• Avec le lancement d'une monnaie virtuelle, le Libra, Facebook va pouvoir collecter davantage de données sur les utilisateurs et ainsi proposer par la suite des crédits et des assurances selon le profil.

• Enfin il est recommandé de ne pas déposer sur le *cloud* des données trop confidentielles, car certaines entreprises s'autorisent à utiliser les données déposées par les utilisateurs.

VII. Protection

• En France, la **CNIL** (Commission Nationale de l'Informatique et des Libertés), créée en 1978, est l'autorité nationale de protection des données personnelles. Elle veille à ce que l'informatique ne porte pas atteinte à la vie privée, à l'identité humaine et aux libertés individuelles. Elle accompagne également les entreprises dans leur mise en conformité vis-à-vis de la loi.

• Le parlement européen a adopté récemment le **RGPD (règlement général sur la protection des données)**. Le texte oblige tout organisme actif dans l'UE qui collecte des données à prouver la nécessité de cette collecte, à protéger les données collectées et à être transparent sur leur utilisation.

• Par exemple en se connectant sur le site web sncf.com, un message est affiché en bas de l'écran : « En poursuivant votre navigation sur ce site, vous acceptez nos CGU et l'utilisation de cookies qui nous permettent de vous proposer une navigation optimale et des contenus adaptés à vos centres d'intérêt. » L'utilisateur doit alors valider en cliquant sur « J'accepte » ou paramétrer les cookies.

• Certaines entreprises se sont vues infliger des amendes pour non-respect du RGPD.