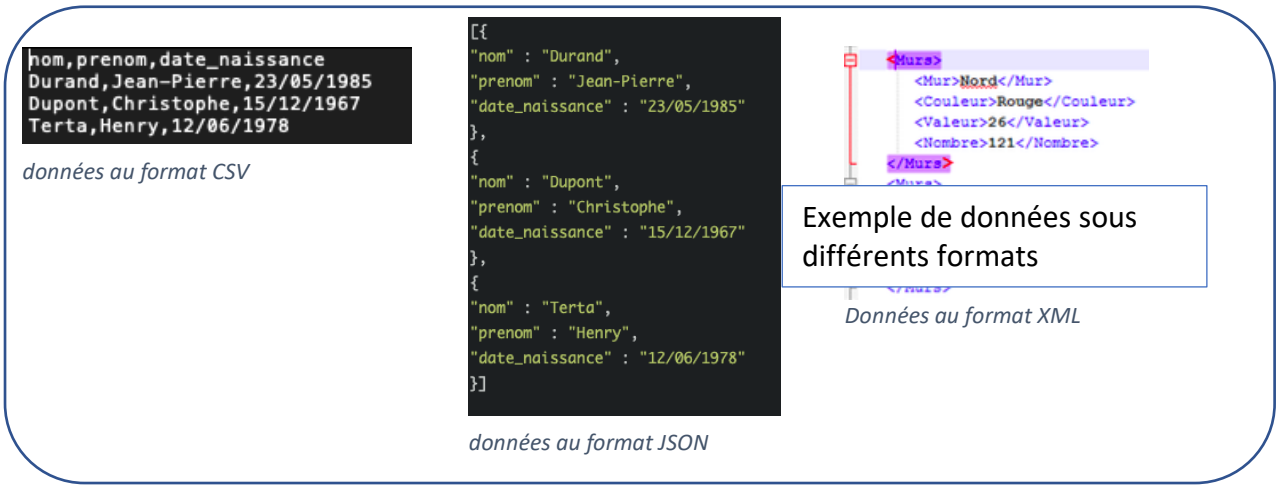
SNT - Thème 3 : Les Données et leurs traitements

Vidéo : Lien

1. Données structurées... Qu'est-ce que c'est ?

Ce sont des informations (mots, chiffres, signes,...) contrôlées par des référentiels et présentés dans des cases (les champs d'une base de données) qui permettent leur interprétation et leur traitement par des machines.
Les données non structurées sont quant à elles les données qui ne sont pas organisées en base de données (ex. : les photos, mails, vidéos que vous recevez dans votre ordinateur).
Il y a des données quantitatifs représentées sous forme de nombres (taille, poids, âge,...) et des données qualitatives sous forme textuelle(nom, matricule,...)



données au format CSV

données au format JSON

Données au format XML

Le format **CSV** (Comma-Separated Values) est un format texte ouvert présentant des valeurs tabulaires sous forme de valeurs séparées par des virgules.

Chaque ligne du texte correspond à une ligne du tableau et les virgules correspondent aux séparations verticales des colonnes.

Le format **JSON** (JavaScript Object Notation – Notation Objet issue de JavaScript) est un format léger d'échange de données. Il est facile à lire ou à écrire pour des humains. Il est aisément analysable ou générable par des machines. Il est basé sur un sous-ensemble du langage de programmation JavaScript.
Le format XML (Extensible Markup Language, généralement appelé XMLnote 1), « langage de balisage extensible1 » en français, est un métalangage informatique de balisage générique qui est un sous-ensemble du Standard Generalized Markup Language (SGML).

2. Vocabulaire :

Descripteur (ou attribut d'une données) : quantité mesurable ou calculable qui permet de décrire en partie un objet, un signal, une donnée.
Ex. : Un contact est décrit par son « nom », son « prénom », son « adresse », son « numéro de téléphone ».
La valeur d'un descripteur est aussi une donnée.

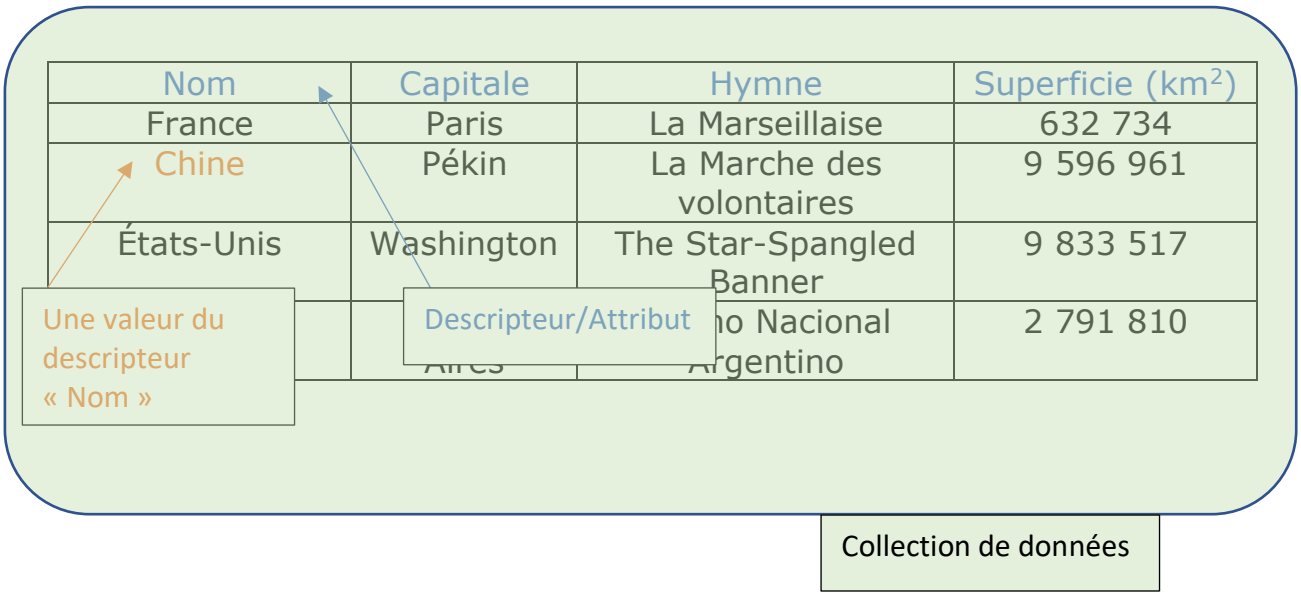
Les descripteurs doivent être choisis de manière pertinente.

Collection de données : Une collection regroupe des objets partageant les mêmes descripteurs (ex. la collection des contacts d'un carnet d'adresses)

Une base de données : Elle regroupe plusieurs collections de données reliées entre elles. (ex. : la base de données d'une bibliothèque regroupe des données sur les livres, les abonnés et les emprunts effectués).

Métadonnée : Donnée qui regroupe d'autres données. Disons plutôt que c'est une information qui permet de retrouver les données stockées (ex. date de sauvegarde, taille et auteur d'un fichier, données GPS d'une photo...). Elles sont très utiles pour les données peu ou pas structurées afin de les rendre accessibles et utilisables.

Open data : Ce sont des données totalement publiques et libres de droits. Lien vers les données publiques françaises : https://www.data.gouv.fr/fr/



3. À quoi ça ressemble ?

À quoi ressemble les différents langage vu plus haut avec cette collection de données :

En CSV :

```
"Nom","Capitale","Hymne","Superficie"
"France","Paris","La Marseillaise","632734"
"Chine","Pékin","La Marche des Volontaires","9596961"
"Etats Unis","Washington","The Star Spangled Banner","9833517"
"Argentine","Bueno Aires","Himno Nacional Argentino","2791810"
```

En JSON :

```
[
   {
      "nom": "France",
      "capitale": "Paris",
      "hymne": "La Marseillaise",
```

```
      "superficie": "632724"
   },
   {
      "nom": "Chine",
      "capitale": "Pékin",
      "hymne": "La Marche des Volontaires",
      "superficie": "9596961"
   },
   {
      "nom": "Etats Unis",
      "capitale": "Washington",
      "hymne": "Star Spangled Banner",
      "superficie": "9833517"
   },
   {
      "nom": "Argentine",
      "capitale": "Bueno Aires",
      "hymne": "Himno Nacional Argentino",
      "superficie": "2791810"
   }
]
```

En XML :

```
<?xml version="1.0" encoding="UTF-8"?>
<repertoire>
      <pays>
            <nom>France</nom>
            <capitale>Paris</capitale>
            <hymne>La Marseillaise</hymne>
            <superficie>632724</superficie>
      </pays>
      <pays>
            <nom>Chine</nom>
            <capitale>Pékin</capitale>
            <hymne>La Marche des Volontaires</hymne>
            <superficie>9596961</superficie>
      </pays>
      <pays>
            <nom>Etats Unis</nom>
            <capitale>Washington</capitale>
            <hymne>Star Spangled Banner</hymne>
            <superficie>9833517</superficie>
      </pays>
      <pays>
            <nom>Argentine</nom>
            <capitale>Bueno Aires</capitale>
            <hymne>Himno Nacional Argentino</hymne>
            <superficie>2791810</superficie>
      </pays>
</repertoire>
```

4. Repères historiques :

- 1930 : utilisation des cartes perforées, premier support de stockage de données.
- 1956 : invention du disque dur permettant de stocker de plus grandes quantités de données, avec un accès de plus en plus rapide.
- 1970 : invention du modèle relationnel (E. L. Codd) pour la structuration et l'indexation des bases de données.
- 1979 : création du premier tableur, VisiCalc.
- 2009 : *Open Government Initiative* du président Obama.
- 2013 : charte du G8 pour l'ouverture des données publiques.

5. Problématiques des données :

- Stockage des données structurées (fiabilité, vélocité, volume, énergie,...)

Chaque jour, 500 térabytes (10^{12} bytes) de données sont stockées sur facebook. En 2012, 43 exabytes (10^{18}) de trafic par mois. 1,8 zettabytes (10^{21}) de données ont été produite en 2011.
10% de la production électrique mondiale est utilisé par le secteur du numérique dont 18% pour les data-centers(lieus où sont stockées les données telles que vos discussions sur messenger ou autres, vos mails, les vidéos Youtube...).
En 2015, la consommation électrique des data centers représentait la consommation électrique d'une ville telle que Lyon.

- Recherche des données (vélocité, fiabilité, indexation,..)
  La base de données de Google contient plusieurs milliers de milliards de liens vers des pages. A chacun de ces liens est rattaché une description de son contenu à l'aide de mots clés.
  Ensuite ces pages sont classées par ordre « d'intérêts ».
- Manipulation des données
- Impacts sociétaux

- Découverte de quelques types de données structurées : lien
- Qu'est-ce que le cloud : lien