

Les données structurées

Introduction

Les données constituent la matière première de toute activité numérique.

Afin de permettre leur réutilisation, il est nécessaire de :

- les conserver de manière persistante.
- Les structurer correctement pour que l'on puisse les exploiter facilement pour produire de l'information.

Repères historiques :

- 1930 : utilisation des cartes perforées, premier support de stockage de données
- 1979 : création du premier tableur, VisiCalc.



Les données - L'information

💡 DÉFINITION :

- Les données sont des ensembles de symboles (mots, nombres, images, sons etc...) pour représenter le monde réel (objets événements etc...).
- Elles peuvent être quantitatives (âge, poids taille, température etc..) ou qualitatives (noms, prénom, adresse etc...)

💡 DÉFINITION :

Une information est une donnée interprétée.

Une information = données + modèle d'interprétation.

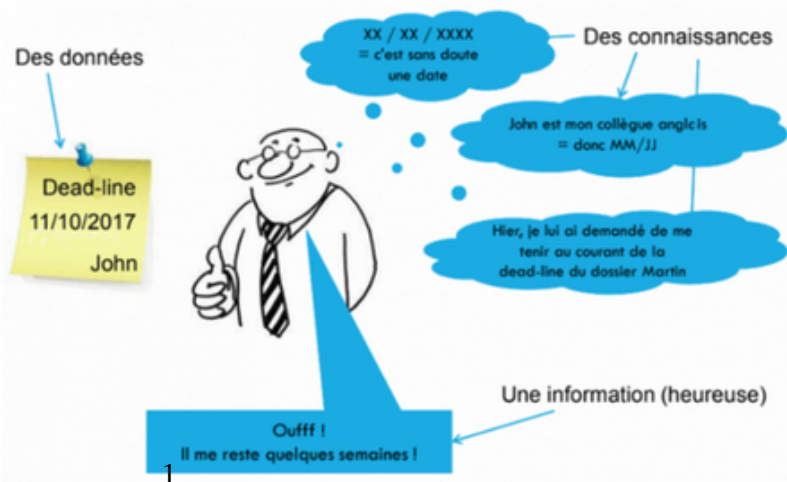
Prenons la donnée suivante : 2 81 12 92 01208680.

C'est juste une série de chiffres...

Si maintenant on précise que c'est un numéro de sécurité sociale, on en déduit qu'il s'agit de celui d'une femme née en 1981 au mois de décembre dans le département 92 (haut de seine).

EXERCICE 1 :

Expliquez ce schéma...



Les données structurées

Les données doivent être décrites, par un **descripteur**, compréhensible pour celui qui veut les interpréter.

Prenons le cas d'une bibliothèque, et considérons un abonné.

Lors de son inscription celui-ci fournit des données (son nom, son prénom, son adresse et son numéro de téléphone), ces données seront associées aux **descripteurs** (Nom, Prénom, Adresse, n-tel).

On regroupe les données des abonnés dans une **table**, avec les mêmes **descripteurs**, on crée ainsi une **collection**.

On regroupe ensuite toutes les **collections** (livres, abonnés, emprunts, etc...) dans une **base de donnée**.

Accès aux données

Certaines données sont accessibles à tous (Open Data - le site data.gouv.fr propose des jeux de données libre d'accès) et d'autres sont non accessibles(privées ou sensibles..).

Certaines bases de données sont souvent comparées à l'or noir d'internet. Les données qu'elles contiennent sont utilisées pour analyser le comportement des internautes afin de leur proposer tel ou tel produit lors de leur navigation.

De puissants algorithmes sont utilisés pour faire ces interprétations.

Représentation des données

Les données sont principalement représentées sous la forme de tableaux. On parle de données tabulaires.

Exemple d'une table listant des fruits vendus par un magasin.

Nom	Prix	Code
Banane	5.99€/kg	77
Pomme	2.99€/kg	99
Poire	7.99€/kg	170

EXERCICE 2 :

Quels sont les descripteurs de la collection ci-dessus?

Il existe trois formats pour représenter un tableau de données : les formats CSV, XML et JSON.

Le format CSV :

Le format Comma-Separated Values (CSV) est un format permettant de représenter des données tabulaires sous la forme de valeurs séparées par des virgules. Le séparateur pourrait être un point-virgule ou tout autre caractère... Le tableau précédent est représenté comme ci-contre :

```
Nom, Prix, Code  
Banane, 5.99, 77  
Pomme, 2.99, 99  
Poire, 7.99, 170
```

Le format XML :

Le format eXtensible Markup Language (XML) est un format qui utilise des balises pour structurer les données dans le fichier texte.

Les balises sont utilisées pour encadrer un contenu: il y a une balise ouvrante et une balise fermante.

```
<data>
  <fruit>
    <name>Banane</name>
    <price>5.99</price>
    <code>77</code>
  </fruit>
  <fruit>
    <name>Pomme</name>
    <price>2.99</price>
    <code>99</code>
  </fruit>
  <fruit>
    <name>Poire</name>
    <price>7.99</price>
    <code>170</code>
  </fruit>
</data>
```

Le format JSON :

Le format JavaScript Object Notation (JSON) est un format plus récent utilisé pour représenter des objets qui dérivent de la notation des objets du langage JavaScript.

On remarquera une structure de dictionnaire qui contient une liste de dictionnaires.

```
{
  "data": [
    {
      "name": "Banane",
      "price": 5.99,
      "code": 77
    },
    {
      "name": "Pomme",
      "price": 2.99,
      "code": 99
    },
    {
      "name": "Poire",
      "price": 7.99,
      "code": 170
    }
  ]
}
```

EXERCICE 3 :

Réaliser "à la main" la représentation de ces données tabulaires dans les trois formats (CSV, XML et JSON)

Nom	Prénom	nombre de voix
Dupond	Émile	514
Dupont	Chloé	632
Dupons	Camille	421