SNT · SECONDE

Données structurées

7 notions, l'essentiel à retenir et les pièges à éviter.

1 Données et métadonnées

Information = donnée + contexte

- Donnée : valeur brute. Information : donnée placée dans un contexte qui lui donne du sens.
- Métadonnée : donnée qui décrit une donnée (date, taille, auteur, position GPS d'une photo).

2 Descripteurs et collections

ligne = objet · colonne = descripteur

- Collection : objets de même nature, décrits par les mêmes descripteurs, présentés en table.
- Chaque descripteur a un type : entier, décimal, texte, date, booléen.
- Identifiant : descripteur dont la valeur est unique pour chaque objet.

3 Le format CSV

- Fichier texte : un en-tête avec les descripteurs, puis une ligne par objet.
- Séparateur : virgule, ou point-virgule quand les nombres ont une virgule décimale.
- En Python, csv.DictReader donne une liste de dictionnaires dont les valeurs sont des chaînes.

4 Rechercher, filtrer, trier

sorted(table, key=cle, reverse=True)

- Filtrer : garder les lignes qui vérifient une condition (for et if en Python, filtre au tableur).
- Trier : ranger les lignes entières selon une clé, en ordre croissant ou décroissant.
- Calculer sur une colonne : somme, moyenne, maximum.

5 Données personnelles et RGPD

- Donnée personnelle : se rapporte à une personne identifiable, directement ou indirectement.
- RGPD, appliqué depuis le 25 mai 2018 ; en France, la CNIL le fait respecter.
- Droits : accès, rectification, effacement, opposition, portabilité, limitation.
- Avant 15 ans, un service fondé sur le consentement doit aussi obtenir celui d'un parent.

6 Volumes et stockage

1 To = 1 000 Go = 1 000 000 Mo

- 1 octet = 8 bits ; ko, Mo, Go, To, Po : chaque unité vaut 1 000 fois la précédente.
- Page de texte : quelques ko ; photo : quelques Mo ; film : quelques Go.
- Cloud : serveurs de centres de données, accessibles par Internet et gourmands en électricité.

7 Données ouvertes et croisement

- Données ouvertes : accessibles à tous, réutilisables librement selon une licence ouverte.
- Croiser deux tables par un descripteur commun crée de nouvelles informations.
- Retirer le nom ne suffit pas à anonymiser : un croisement peut réidentifier.

Pièges à éviter

- Confondre le descripteur (le nom de la colonne) et la valeur (le contenu de la case).
- Compter la ligne d'en-tête d'un CSV comme un objet.
- Calculer avec des valeurs lues dans un CSV sans les convertir : ce sont des chaînes.
- Trier une seule colonne au tableur : les valeurs ne correspondent plus aux bons objets.
- Croire qu'une donnée sans nom est forcément anonyme.