CH 1 - Les données structurées et leur traitement

1. Les données

1.1. Les données

Les données constituent la matière première de toute activité numérique.

Une donnée est un élément se rapportant à un objet, une personne ou un événement.

Une donnée personnelle est une donnée identifiant directement ou indirectement une personne physique (Nom, numéro de téléphone, photographie, date de naissance, empreinte digitale, etc.).

1.2. Les métadonnées

Une métadonnée est une donnée particulière qui apporte des informations sur la donnée principale.

Exemples :

- Pour un fichier de musique, la donnée principale est la musique elle-même ; le nom de l'artiste ou le nom de l'album sont des métadonnées.
- Pour une photo, la donnée principale est l'image ; le nom de la photo ou la date de la prise de vue sont des métadonnées.

2. Les données structurées

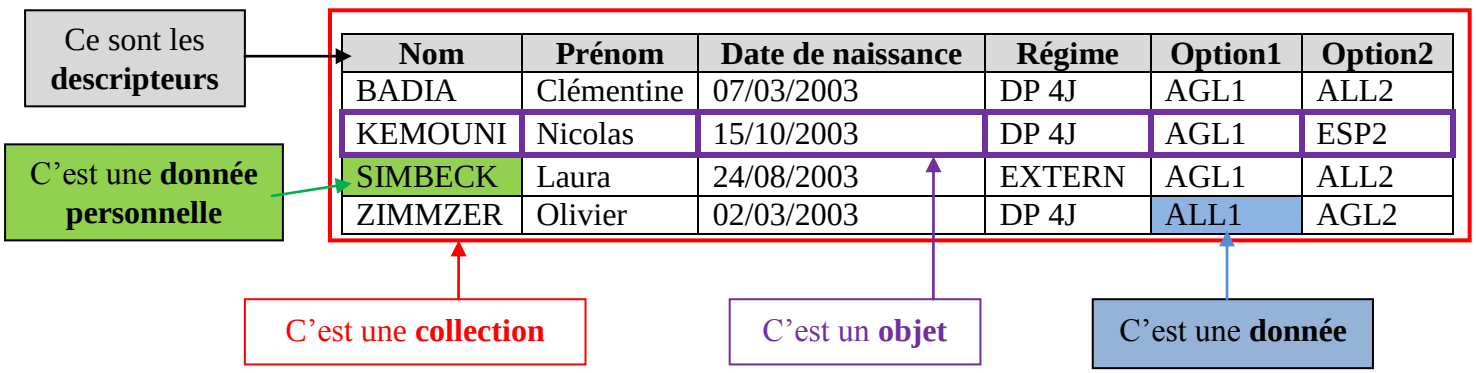
Afin de permettre leur réutilisation, il est nécessaire de conserver les données de manière persistante. Les **structurer** correctement garantit que l'on puisse les exploiter facilement pour produire de l'information.

Remarque : les données non structurées peuvent aussi être exploitées, par exemple par les moteurs de recherche.

2.1. Descripteurs et collection

Un « objet » peut comporter plusieurs données. Il faut donc un titre pour décrire la donnée, on le nomme un **descripteur**. Plusieurs descripteurs peuvent donc être utiles pour décrire un même objet.
Si plusieurs « objets » sont décrits avec les mêmes descripteurs on les regroupe dans un tableau que l'on nomme **collection**.

Exemple :
Pour l'organisation du lycée il existe des tableaux regroupant tous les élèves, leurs données personnelles, leur régime (DP = demi pensionnaire), leurs options (AGL = anglais) …



Nom	Prénom	Date de naissance	Régime	Option1	Option2
BADIA	Clémentine	07/03/2003	DP 4J	AGL1	ALL2
KEMOUNI	Nicolas	15/10/2003	DP 4J	AGL1	ESP2
SIMBECK	Laura	24/08/2003	EXTERN	AGL1	ALL2
ZIMMZER	Olivier	02/03/2003	DP 4J	ALL1	AGL2

2.2. Bases de données

Une **base de données** regroupe plusieurs collections de données reliées entre elles. Par exemple la base de données du lycée conserve les données des élèves, des professeurs et des autres personnels (agents, direction …).

3. Le traitement des données

3.1. Les formats

Pour assurer la persistance des données, ces dernières sont stockées dans des fichiers.

Pour enregistrer une table le format **CSV** (**C**omma **S**eparated **V**alues, données séparées par des virgules) est un format de fichier simple très utilisé. D'autres formats existent comme JSON pour JavaScript Object Notation ou XML pour eXtensible Markup Language.

Pour d'autres fichiers les formats sont standardisés en fonction de l'application :

- pour des sons : MP3, WAV, …
- pour des textes : TXT, DOC, ODT, …
- pour des tableurs : XLS, ODS, …
- pour des vidéos : AVI, MPEG, …
- pour des images : BMP, JPG, …

3.2. Les traitements

Une table de données peut faire l'objet de différentes opérations :

- rechercher une information précise dans la collection,
- trier la collection sur une ou plusieurs propriétés,
- filtrer la collection selon un ou plusieurs tests sur les valeurs des descripteurs,
- effectuer des calculs,
- mettre en forme les informations produites pour une visualisation par les utilisateurs.

La recherche dans une base comportant plusieurs collections peut aussi croiser des collections différentes sur un descripteur commun ou comparable.

4. Le cloud

4.1. Le stockage

Les fichiers de données sont stockés sur des supports de stockage : internes (disque dur ou SSD) ou externes (disque, clé USB), locaux ou distants (cloud). Ces supports pouvant subir des dommages entraînant des altérations ou des destructions des données, il est nécessaire de réaliser des sauvegardes.

4.2. L'impact sociétal et environnemental

Les grandes bases de données sont souvent implémentées sur des serveurs dédiés (machines puissantes avec une importante capacité de stockage sur disques). Ces centres de données doivent être alimentés en électricité et maintenus à des températures suffisamment basses pour fonctionner correctement.

L'exploitation de données massives (Big Data) est en plein essor dans des domaines aussi variés que les sciences, la santé ou encore l'économie. Les conséquences sociétales sont nombreuses tant en termes de démocratie, de surveillance de masse ou encore d'exploitation des données personnelles.

Certaines de ces données sont dites ouvertes (OpenData), leurs producteurs considérant qu'il s'agit d'un bien commun. Mais on assiste aussi au développement d'un marché de la donnée où des entreprises collectent et revendent des données sans transparence pour les usagers. D'où l'importance d'un cadre juridique permettant de protéger les usagers, préoccupation à laquelle répond le règlement général sur la protection des données (RGPD).

Les centres de données (datacenter) stockent des serveurs mettant à disposition les données et des applications les exploitant. Leur fonctionnement nécessite des ressources (en eau pour le refroidissement des machines, en électricité pour leur fonctionnement, en métaux rares pour leur fabrication) et génère de la pollution (manipulation de substances dangereuses lors de la fabrication, de la destruction ou du recyclage). De ce fait, les usages numériques doivent être pensés de façon à limiter la transformation des écosystèmes (notamment le réchauffement climatique) et à protéger la santé humaine.