4 Les données structurées et leur traitement

Une donnée est la représentation d'une information. Bien avant la naissance de l'informatique, on a eu besoin de stocker, puis de structurer ces informations, pour pouvoir les utiliser.
Les données constituent désormais la matière première de toute activité numérique. À ce titre, de nouvelles formes de stockage, de structuration et d'exploitation ont vu le jour au regard de l'explosion de la quantité de données disponibles, jusqu'à l'émergence d'une science du traitement des données : la **data science**.

- Repères historiques

1928 : cartes perforées 80 colonnes de IBM
1956 : invention du disque dur
1979 : VisiCalc premier tableur,
2000 : émergence de la data science
2013 : charte du G8 pour l'ouverture des données publiques.

1 Structuration des données

Une **donnée** est une valeur décrivant un objet. Par exemple, le numéro de téléphone d'un contact est une donnée. Plusieurs **descripteurs** peuvent être utiles pour décrire un objet. Par exemple, pour caractériser un contact : nom, prénom, adresse, numéro de téléphone).
Une **collection** regroupe des objets partageant les mêmes descripteurs (par exemple, la collection des contacts d'un carnet d'adresses).
Une **base de données** regroupe plusieurs collections de données reliées entre elles. Par exemple, la base de données d'une bibliothèque conserve les données sur les livres, les abonnés et les emprunts effectués.
Comme sur papier, on utilise souvent des **tableaux** pour organiser les données numériques en colonnes et en lignes. En informatique, on appelle **table**, un tableau dans lequel la première ligne sert à décrire la forme des lignes suivantes et on appelle **nom de champ** (ou critère) l'intitulé qui sert à décrire la nature des informations notées dans les colonnes.
Dans la table ci-contre, présentant des données sur la France et ses pays limitrophes, les noms de champs sont ***Pays****,* ***Population*** *et* ***Superficie***

Pays	Population (janv 2019)	superficie (en km²)
France	66 992 699	551 695
Allemagne	82 886 977	357 386
Belgique	11 376 070	30 688
Espagne	46 698 569	505 992
Italie	60 494 785	302 072
Luxembourg	613 894	2 586
Royaume-Uni	66 465 641	242 545
Suisse	8 542 323	41 285

2 Stockage des données

▶ **Format :** Pour mémoriser les tables dans un ordinateur, on les stocke dans des fichiers à différents formats dont le rôle est de préciser comment sont organisées les différentes lignes et colonnes.
Plusieurs formats sont couramment utilisés parmi lesquels :

- Le format **CSV** (*Comma Separated Values*) dans lequel chaque ligne contient des valeurs séparées par des symboles de ponctuations. La première ligne contenant les noms des champs.
- Le format **ODS** (*Open Document Spreadsheet*), utilisé par les logiciels tableur (du type : Excel) .
- Le format **JSON** (JavaScript Object Notation), format de données textuelles en paires de nom/valeur. Les nom et valeur sont séparés par deux points « : » et chaque paire est séparée de la suivante par une virgule.

▶ **Métadonnées :** À tout fichier sont associées des **métadonnées** qui permettent d'en décrire le contenu. Ces métadonnées varient selon le type de fichier (date et coordonnées de géolocalisation d'une photographie, auteur et titre d'un fichier texte, etc.)
▶ **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***). Aujourd'hui, des centres de données (Data centers) hébergent un nombre toujours croissants de données mais posent des problèmes écologiques : consommation d'énergie en hausse pour fonctionnement des serveurs et des climatisations nécessaires.
▶ **Propriétés des données :** Certaines des données sont dites ouvertes (OpenData) et permettent des usages

libres. Mais on assiste aussi au développement d'un marché de collecte et de vente de données par des entreprises spécialisées, parfois sans informer les usagers. La France a donc choisi de mettre en place un cadre juridique permettant de protéger les usagers : le règlement général sur la protection des données (RGPD).

3 Traitement des données

▶ Les logiciels **tableur** sont des outils pour traiter des données organisées en colonnes et en lignes. Ils permettent de :

- **trier des données** d'une table (modifier l'ordre des lignes selon un descripteur choisi),
- **filtrer des données** d'une table (sélectionner les données contenant une information particulière),
- **effectuer des calculs,**
- **mettre en forme des données** d'une table pour une meilleure visualisation (représentation graphiquement des données).

▶ Pour effectuer un traitement particulier, ou pour l'automatiser, on peut aussi la programmer. Python est un **langage de programmation** adapté au traitement de données.

▶ Aujourd'hui les algorithmes sont capables de traiter un grand nombre de données. L'exploitation de ces données massives (Big Data) permettent d'ouvrir des horizons différents dans le domaine des sciences, de la santé, de l'économie mais posent aussi question sur les impacts relatifs à la démocratie et à la protection des libertés individuelles.

4 Impact sur les pratiques humaines

▶ L'évolution des capacités de stockage, de traitement et de diffusion des données fait qu'on assiste aujourd'hui à un phénomène de surabondance des données et au développement de nouveaux algorithmes capables de les exploiter.
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.