LES DONNEES STRUCTUREES ET LEUR TRAITEMENT
BILAN (ce que dit le programme)

- **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 garantit que l'on puisse les exploiter facilement pour produire de l'information. Cependant, les données non structurées peuvent aussi être exploitées, par exemple par les moteurs de recherche.

- **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.

- **Les données et l'information**

Une **donnée** est une valeur décrivant un objet, une personne, un événement digne d'intérêt pour celui qui choisit de la conserver. Par exemple, le numéro de téléphone d'un contact est une donnée. Plusieurs **descripteurs** peuvent être utiles pour décrire un même objet (par exemple des descripteurs permettant de caractériser un contact : nom, prénom, adresse et numéro de téléphone).

Une **collection** regroupe des objets partageant les mêmes descripteurs (par exemple, la collection des contacts d'un carnet d'adresses). La structure de table permet de présenter une collection : les objets en ligne, les descripteurs en colonne et les données à l'intersection. Les données sont alors dites structurées.

Pour assurer la persistance des données, ces dernières sont stockées dans des fichiers. Le format CSV (*Comma Separated Values*, les données avec des séparateurs) est un format de fichier simple permettant d'enregistrer une table. 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.).

Les données comme les métadonnées peuvent être capturées et enregistrées par un dispositif matériel ou bien renseignées par un humain. Elles sont de différents types (numériques, textes, dates) et peuvent être traitées différemment (calcul, tri, affichage, etc.).

Certaines collections typiques sont utilisées dans des applications et des formats standardisés leur sont associés : par exemple le format ouvert vCard (extension .vfc) pour une collection de contacts.

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.

- **Les algorithmes et les programmes**

La recherche dans des **données structurées** a d'abord été effectuée selon une indexation préalable faite par l'homme. Des algorithmes ont ensuite permis d'automatiser l'indexation à partir de textes, d'images ou de sons.

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.

- **Les machines**

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.

Des recherches dans les fichiers se font à l'intérieur même des ordinateurs, soit sur la base de leurs métadonnées, soit sur la base d'une indexation (à la manière des moteurs de recherche sur le *Web*).

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.

- **Impacts 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.

Contenus	Capacités attendues
Données	Définir une donnée personnelle. Identifier les principaux formats et représentations de données.
Données structurées	Identifier les différents descripteurs d'un objet. Distinguer la valeur d'une donnée de son descripteur. Utiliser un site de données ouvertes, pour sélectionner et récupérer des données.
Traitement de données structurées	Réaliser des opérations de recherche, filtre, tri ou calcul sur une ou plusieurs tables.
Métadonnées	Retrouver les métadonnées d'un fichier personnel.
Données dans le nuage (*cloud*)	Utiliser un support de stockage dans le nuage. Partager des fichiers, paramétrer des modes de synchronisation. Identifier les principales causes de la consommation énergétique des centres de données ainsi que leur ordre de grandeur.

Exemples d'activités

- Consulter les métadonnées de fichiers correspondant à des informations différentes et repérer celles collectées par un dispositif et celles renseignées par l'utilisateur.
- Télécharger des données ouvertes (sous forme d'un fichier au format CSV avec les métadonnées associées), observer les différences de traitements possibles selon le logiciel choisi pour lire le fichier : programme Python, tableur, éditeur de textes ou encore outils spécialisés en ligne.
- Explorer les données d'un fichier CSV à l'aide d'opérations de tri et de filtre, effectuer des calculs sur ces données, réaliser une visualisation graphique des données.
- À partir de deux tables de données ayant en commun un descripteur, montrer l'intérêt des deux tables pour éviter les redondances et les anomalies d'insertion et de suppression, réaliser un croisement des données permettant d'obtenir une nouvelle information.
- Illustrer, par des exemples simples, la consommation énergétique induite par le traitement et le stockage des données.