# LES DONNEES ET LEUR TRAITEMENT

## I- Généralité :

L'information humaine est transcrite sous forme de données pour être manipulée numériquement.
On distingue les données qui doivent être entrées dans la machine, des résultats de calculs, ou sortie des algorithmes.
Les données en tant qu'objets numériques forment un bien non rival* dont la copie ne coûte quasiment rien, et que l'on peut dupliquer sans le consommer.
La production gigantesque de données pose des problèmes planétaires en matière d'environnement (consommation énergétique, utilisation de ressources naturelles rares).
La prolifération de données pose aussi un problème de pérennité à long terme (plusieurs dizaines d'années) qui n'est pas résolu actuellement.
(*) bien non rival : la plupart des objets matériels sont des biens rivaux c'est-à-dire que si on les consomme ou les utilise, ils ne sont plus disponibles pour les autres consommateurs, ce n'est pas le cas des objets informationnels comme par exemple une bonne histoire : si je la partage, elle reste intacte, voire elle s'enrichit.

### A- Définition :

**1-** Mégadonnées : (bigdata) on parle de mégadonnée quand le volume de données est tel qu'on peut faire des analyses statistiques qui permettent de prédire des informations, même si les données sont très diverses, sans information structurée, et approximatives.

2- Données ouvertes : (opendata) données numériques, d'origine publique ou privée, diffusées de manière structurée selon une méthode et une licence libre, garantissant leur libre accès et leur réutilisation par toutes et tous, sans restriction technique, juridique ou financière.

3- Licence libre : licence s'appliquant à une œuvre de l'esprit (document, logiciel, …) par laquelle l'autrice ou l'auteur concède les droits que lui confère le droit d'auteur : usage de l'œuvre, étude de l'œuvre pour en comprendre le fonctionnement ou l'adapter à ses besoins, modification (amélioration, extension et transformation) ou incorporation de l'œuvre en une œuvre dérivée, redistribution de l'œuvre, c'est-à-dire sa diffusion à d'autres usagers, y compris commercialement.

4- Informatique durable : (green IT) vise à réduire l'empreinte écologique, économique et sociale des technologies de l'information et de la communication.

5- Règlement général sur la protection des données (RGPD) : renforce et unifie la protection des données pour les personnes au sein de l'Union Européenne.

## B- Contenu :

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.

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 (Open Data), 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.

## *C- Historique :*

L'idée de pouvoir traiter mécaniquement de l'information est ancienne, dès le XVIIème siècle, par exemple, Gottfried Wilhelm Leibniz va chercher à établir une langue dite caractéristique universelle, qui permettrait d'exprimer la totalité des pensées humaines et pourrait résoudre des problèmes par un calculateur (*calculus ratiocinator*), anticipant l'informatiquede plus de trois siècles. Il faudra attendre le XXème siècle pour comprendre qu'une telle machine est un objet impossible ne serait-ce qu'en mathématiques, d'après les théorèmes d'Alonzo Church et Alan Turing : très simplement, certains calculs (par exemple savoir si un programme va boucler à l'infini, le problème de l'arrêt) nécessitent des temps … infinis. On commençait à comprendre les limites de l'intelligence mécanique avant même de l'avoir fabriquée.

Ces mêmes personnes ont pourtant fondé, dans les années 1930, l'informatique, un domaine d'activité scientifique, technique et industriel concernant le traitement automatique de l'information par l'exécution de programmes informatiques par des machines. On peut attribuer à Ada Lovelace, un siècle avant, d'avoir compris que l'on peut «  calculer sur des nombres mais aussi sur des symboles », on parlerait de données numériques et symboliques aujourd'hui.  Il est intéressant de noter que ces idées sont nées avant la technologie permettant de les mettre en œuvre.

parallèlement, dans les années 1880, Herman Hollerith, futur fondateur d'IBM, fonde la mécanographie en inventant une machine électromécanique destinée à faciliter le recensement en stockant les informations sur une carte perforée. Ces premières cartes perforées ont fait leur apparition au XVIII$^{e}$ siècle dans divers automates et en particulier les métiers à tisser, les orgues de Barbarie et les pianos mécaniques.

L'histoire de l'informatique va véritablement commencer au milieu du XXème siècle avec l'architecture de von Neumann, mise en application de la machine universelle de Turing : les ordinateurs dépassent la simple faculté de calculer et peuvent commencer à traiter des données …

- 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.

## D- Représentation d'une information :

Toutes les informations humaines se codent en binaire ; bien entendu ce n'est pas l'objet réel, ce n'est que son reflet numérique.
Une donnée est spécifiée par des valeurs et chaque valeur a un type (par exemple : vrai ou faux, on dit booléen ; ou bien numérique ou textuel ; ou encore un type spécifique, comme une date) ; selon le type de la donnée on ne fait pas les mêmes opérations.
Une donnée se décompose de manière atomique en données élémentaires, par exemple le nom d'une personne en prénom et patronyme. Bien structurer les données facilite leur traitement par des algorithmes.
Une collection de données peut être ordonnée sous forme de liste, ou bien sans ordre sous forme d'un ensemble.
La façon de structurer les données influe fortement sur les opérations de traitement : il est par exemple bien plus efficace de rechercher une donnée dans une collection toujours ordonnée, mais y insérer une information est plus coûteux.

Donnée : représentation d'une information au sein d'un système informatique.
Métadonnée : donnée servant à définir ou décrire une autre donnée, pour permettre sa manipulation.
Une base de données regroupe plusieurs collections de données reliées entre elles.
Descripteur : mot ou un groupe de mots choisi pour caractériser les informations contenues dans un document et pour faciliter les recherches.

### 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.

## E- 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 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.