Les données structurées et leur traitement

I) Généralités

1) Définition, problématique

Une **donnée** est une valeur décrivant un objet, une personne, un événement digne d'intérêt pour celui qui choisit de la conserver. **Tout peut être potentiellement une donnée**, votre température, votre année de naissance, la race de votre chien.

Comme on a pu le voir avec les moteurs de recherches et les index gigantesques qui permettent de faire le tri dans les pages, le web qui multiplie les textes, les photos et les vidéos, ou encore les réseaux sociaux dans lesquels les usagers partages des données, les données brassées sont de plus en plus nombreuses.

Toutes les données informatiques quelles qu'elles soient sont converties en binaire, c'est-à-dire en 0 et en 1, si bien qu'il n'y a pas de différence pour le stockage entre une photo, un texte ou une vidéo. Se pose le problème des espaces de stockage : le support sur lequel on conserve les données bien sûr, les lieux de stockage, mais aussi le coût énergétique que cela représente.

Les données brutes n'ont aucun sens sans traitement. Concrètement relever les températures de tous les individus que vous croisez dans la rue n'a pas de sens si vous n'exploitez pas derrière. Si par contre, dans le cadre d'une épidémie vous voulez lier maladie et température des passants, alors vous avez besoin d'un logiciel qui permet d'étudier ces données et de représentations. Le tableur, les bases de données, sont autant d'outils qui permettent de traiter les données. Le traitement des données implique un accès à ces données, données qui peuvent être sensibles, votre religion, votre sexualité, vos informations bancaires. Il faut alors faire la distinction entre une donnée **personnelle** qui permet d'identifier une personne physique et une donnée publique. La RGPD est une loi Européenne qui vise à protéger les données des concitoyens. En France c'est la CNIL (Commission Nationale de l'Informatique et des Libertés) qui veille au respect de la loi.

2) Historique

2400 avant J.-C., l'invention de l'écriture entraîne les premiers stockages de comptabilité agricole sur des tablettes en argile, c'est le début du traitement des données.
1930 : utilisation des cartes perforées, premier support de stockage de données
1950 : utilisation des bandes magnétiques pour le stockage, elles sont encore très largement utilisées aujourd'hui.
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. Même année, création du CD-ROM
1997 : apparition du terme Big Data
1999 : première clé USB
2000 : premiers SSD, qui remplace de plus en plus le disque dur pour sa grande rapidité
2003 : premier Bluray
2013 : charte du G8 pour l'ouverture des données publiques.
2037 : année où la production d'électricité sera insuffisante pour alimenter l'ensemble des ordinateurs de la planète

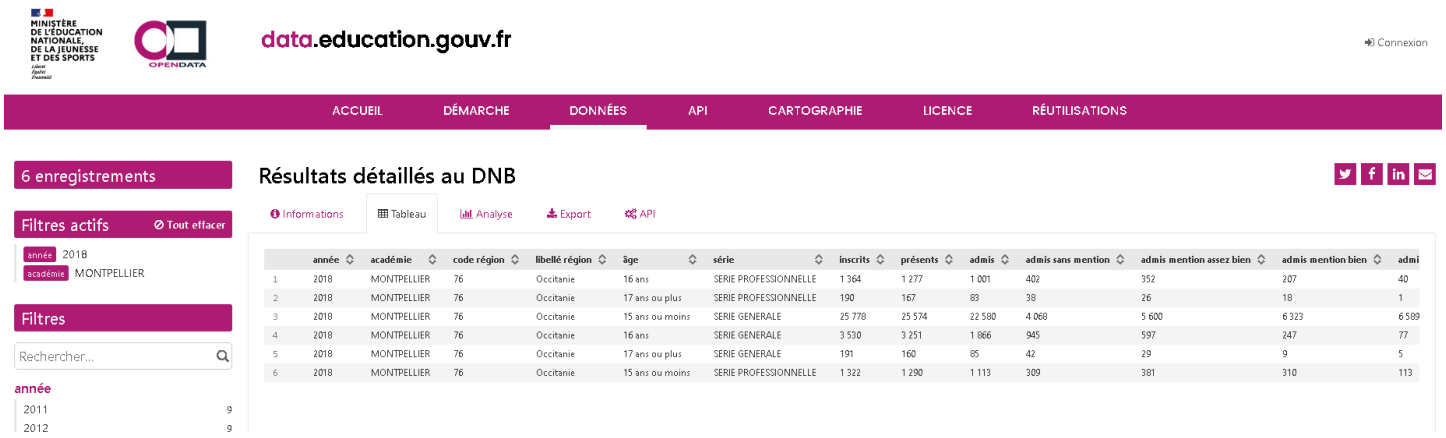
II) Données structurées

1) Dans des tableaux

Lorsque les volumes de données sont importants, on a besoin de **descripteurs** pour réussir à les trier. Si on prend un ensemble de personnes, le numéro de téléphone, l'année de naissance, l'adresse, sont autant de descripteurs disponibles.

On présente les données dans des tableaux ou des tables, c'est une manière visuelle, facile de trier les données.

Activité : Aller sur le site https://data.education.gouv.fr/pages/accueil/ aller sur Résultats détaillés au DNB, faire un choix de l'année 2018, plus choisir l'académie de Montpellier. Faire une sélection sur des valeurs précises des données s'appelle **filtrer**. Donner cinq descripteurs de votre choix qui permettent de trier ces données. Dans les tableaux, on appelle ces descripteurs des **champs**.



Remarque : dans le cas d'élèves comme c'est le cas ici, le nom, le prénom, et même la date de naissance ne sont pas des éléments permettant de distinguer sans ambiguïté deux enfants. En effet, il est tout à fait possible que des élèves portent le même nom et soient nés le même jour. Afin d'avoir la certitude dans le traitement des données de ne pas avoir de confusion possible, l'un des champs est une clé primaire, c'est-à-dire un identifiant unique. À l'école c'est l'INE, identifiant national élève, pour les adultes c'est souvent le numéro de sécurité social.

Activité : à l'aide du tableur, réaliser un tableau avec pour champs, le nom, le prénom, l'âge de chaque élève. Faire un diagramme circulaire pour représenter la situation.

2) Dans les fichiers au format CSV

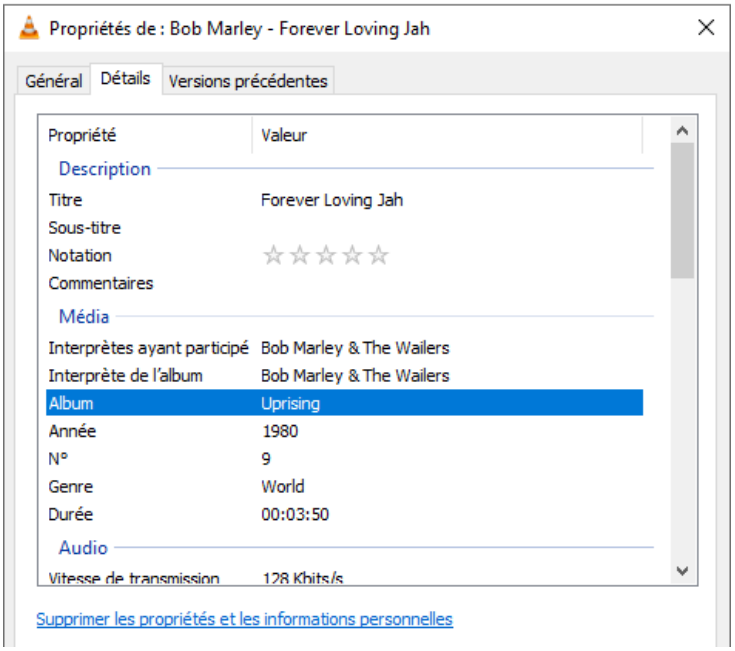
Avec un simple éditeur de texte comme notepad vous pouvez réaliser un fichier de données structurées. Il vous suffit d'appliquer la présentation suivante.

Dans la première ligne descripteur1;descripteur2;descripteur3...descripteurX. Les lignes suivantes doivent avoir la même structure, on veillera à ne pas réutiliser le point virgule dans une des données sinon ce serait interprété comme un descripteur de plus. Selon les pays, le descripteur peut être la virgule ou le point virgule.

Activité : Créer le fichier CSV avec pour descripteur, le nom de l'élève, le prénom de l'élève, le village où il réside. Ouvrir le fichier dans le tableur et vérifier

3) Dans tous les fichiers sous formes de métadonnées

Les **métadonnées** sont des données dans les données. Par exemple un clic droit sur un fichier mp3 permet d'obtenir des informations supplémentaires comme le nom de l'album, l'année de la création ou encore le style musical. Ces informations se retrouvent pour de nombreux fichiers, elles ne seront pas identiques d'un type de fichier à l'autre. Pour une photo par exemple, on obtiendrait la marque de l'appareil ou les coordonnées GPS du lieu où a été pris la photo.



III) Le cloud et ses enjeux

1) Consommation électrique

Cloud en anglais veut dire nuage, c'est un terme mal employé pour définir l'informatique distante. Le cloud c'est ce qui ne se trouve pas sur votre ordinateur. Si vous travaillez dans le lycée sur un fichier qui est sur le serveur de l'école, vous êtes dans un système de cloud. **Le cloud c'est l'ordinateur de quelqu'un d'autre.** Le cloud n'est pas récent, il correspond aux débuts du web, vos mails sont hébergés sur un serveur distant, les pages web sont hébergés sur des serveurs web, la nouveauté aujourd'hui c'est l'hébergement massif de fichiers mais aussi d'applications. Vous pouvez très bien faire vos retouches photos, taper du texte ou même utiliser un ordinateur à distance. **Toutes ces opérations nécessitent des infrastructures de plus en plus puissantes qui vont consommer des quantités d'électricité de plus en plus grandes et contribuer au réchauffement climatique. Chaque action sur internet a un coût énergétique considérable, 20 % de la consommation électrique mondiale provient des datacenters**

D'après https://www.bfmtv.com/sciences/la-consommation-electrique-cachee-de-vos-activites-sur-internet_AN-201805250002.html

2) Révolution technologique ou business ?
Le cloud computing présente de nombreux avantages qui ont été mis en avant pour inciter le consommateur à l'utiliser. Si certains sont vrais, d'autres sont largement plus discutables :

- Avant vous achetiez un DVD et vous le lisiez directement sur votre ordinateur ou votre lecteur DVD. Avec le streaming on fait l'économie d'un objet on polluerait donc moins la planète. Pourtant on voit que le streaming consomme énormément d'électricité, on ne tient pas compte non plus du fait de revoir le film. Regarder votre DVD cinquante fois ne coûte rien ou presque en électricité, ce n'est pas du tout le cas pour le streaming et l'explication se trouve pourtant ici. Quelqu'un qui achète un objet en est le propriétaire. Il peut donc utiliser son objet autant de fois qu'il le veut et le conserver tant qu'il fonctionne et donne satisfaction. Sur une plateforme de streaming, vous n'êtes plus propriétaire de l'objet, ou du moins vous en êtes propriétaire tant que vous payez l'abonnement. Nous sommes passés de l'époque de la propriété à l'époque de l'abonnement. Quand le consommateur paye une seule fois, il paye désormais tout le temps, assurant une rente au service. On notera que cette tendance a commencé par la musique car les fichiers étaient légers, aujourd'hui la vidéo et demain c'est le jeu vidéo qui se joueront en streaming. Vous ne serez ni propriétaire de vos musiques, de vos films ou de vos jeux, vous vous contenterez de les louer.

- Avant vos données étaient présentes sur votre disque dur. Malheureusement en cas de crash de la machine ou de disparition de la machine (vol, perte, catastrophe naturelle), vous perdiez l'intégralité de vos précieuses données si vous n'aviez pas fait de copie de sauvegarde. Avec des services comme dropbox, onedrive ou Gdrive plus de risques de perte de vos données puisqu'elles ne sont plus physiquement sur votre ordinateur, et en plus elles sont accessibles partout. Sauf que si ces services sont gratuits, ils ne le sont que pour 5 Go de données ce qui est très peu. On pourra donc sauvegarder des documents essentiels, mais impossible d'y mettre l'intégralité de ses photos, ses vidéos ou ses musiques sauf si on paye un abonnement. De la même manière que pour le streaming, inciter au cloud c'est inciter à s'abonner donc à payer de façon régulière.

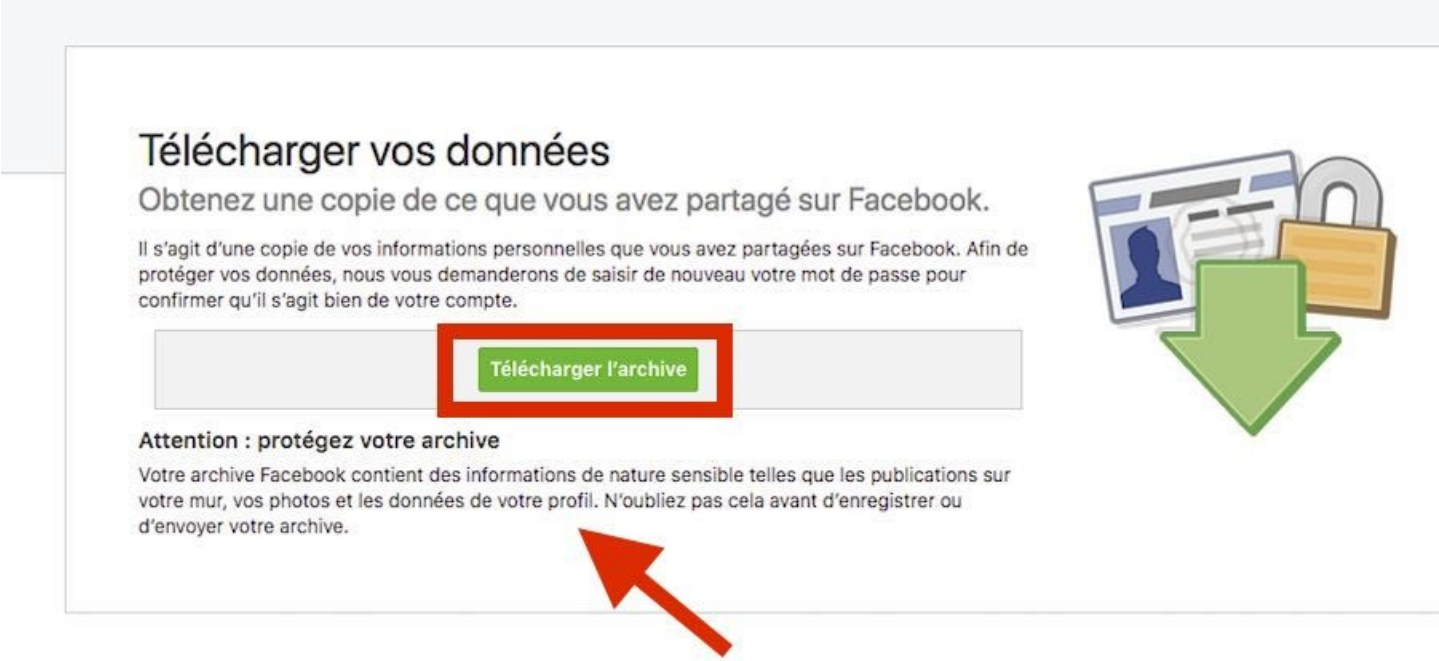
- Avec les services dans le cloud, vos programmes sont toujours à jour, ils bénéficient des dernières nouveautés et sont sécurisés. Il s'agit certainement du point le moins discutable même s'il découle à nouveau sur une obligation d'abonnement au service pour pouvoir utiliser un logiciel. Par exemple, quelqu'un qui a acheté Word 2007 et qui en est parfaitement content peut continuer à l'utiliser sans aucun problème, comme pour un DVD il l'a payé une fois et peut s'en servir à volonté. Néanmoins, est-il en droit de demander des améliorations ou de signaler des problèmes avec les fichiers actuels, quand le programme a plus de 15 ans. L'une des grosses difficultés pour les développeurs, c'est d'avoir des questions sur des versions qui ne sont pas à jour, car même sur des logiciels gratuits, les gens ne font pas les mises à jour. Si les gens se connectent à un service, tout le monde a le même service, tout le monde voit la même chose ce qui facilite largement le travail développeurs.

Ce qu'il faut retenir c'est que le cloud correspond à une tendance forte de notre société, une dématérialisation où l'objet disparaît au profit de services de location.

3) Big data
Big data ou « grosses données » est une appellation pour désigner le très grand nombre de données en circulation. Certains avantages sont évidents. Si on combine des millions de données récoltées sur un thème précis et qu'on fait travailler une intelligence artificielle, on peut alors trouver des solutions pour lutter contre des maladies par exemple. La récupération des données peut avoir du sens, mais c'est loin d'être le cas. La Chine par exemple récolte de très nombreuses données sur ses concitoyens à des fins d'espionnage. Grace à un système de récupération des données très poussé, la Chine a mis en place un système d'échelle sociale donnant des malus aux individus ne respectant pas les codes très stricts de la société. Être en bas de l'échelle, c'est ne plus pouvoir prendre les transports en commun.

La collecte de la data ne peut se faire sans **transparence**, et sans protection. C'est ainsi que vous avez le droit de récupérer l'intégralité de vos données sur les différents réseaux sociaux. Un journaliste avait fait l'expérience il y a quelques années avant que les données ne soient accessibles, on lui avait envoyé une centaine de pages retraçant toute son activité.

Activité : qui est Edward Snowden ? Qu'a-t-il fait ?



Récupération des données sur Facebook

La dernière question qu'on peut se poser quant au big data c'est de savoir si nous avons réellement besoin de tout ça. On parle aujourd'hui d'infobésité tellement nous nous abreuvons de contenus pertinents ou non. Nos cerveaux n'arrivent plus à suivre et pourtant on nous en donne de plus en plus. De la même manière que nous devrions nous interroger quant à la pertinence d'être sur cinq réseaux sociaux, il faut s'interroger sur la quantité d'informations, de séries télés, de films, de vidéos que nous ingérons au quotidien. Est-ce réellement indispensable ?

Version 1.0 : le 12/07/2020 premier jet.