Chapitre 2

Les données et leurs traitements

Contenus	Capacités attendues
Données	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.

En informatique, une donnée est l'enregistrement d'une information. Les données peuvent être conservées et classées sous différentes formes : textuelles (chaîne), numériques, images, sons, etc. Le terme traitement de données renvoie à une série de processus qui permettent d'extraire de l'information ou de produire du savoir à partir de données brutes. Par exemple, je peux enregistrer dans un fichier, le nom de chaque élève d'une classe et y associer leur nombre de frère. Ce fichier comporte des données brutes. Je peux ensuite l'analyser et calculer le nombre moyen de frères et soeurs par élève : j'aurai alors réaliser un traitement des données.

1 Les différents types de données

La forme et les moyens de diffusion des données n'ont pas cessé d'évoluer au cours des siècles : depuis les premières tablettes d'argiles de Mésopotamie 2400 ans avant J.C, les papyrus d'Égypte, les premiers livres recopiés par des générations successives de moines, la diffusion des savoirs grâce à l'invention de l'imprimerie en 1450 par Gutemberg, les premières données informatiques stockées sur cartes perforées en 1928, nous voila à l'aire du Cloud : des données stockées sous forme de bit sur des serveurs situés aux quatres coins de la planètes et accessibles à tous et toutes.

Le format des données informatiques est spécifié par une extension de nom de fichier est un suffixe de nom de fichier fait pour identifier son format. Par exemple un fichier *bla.txt* est un fichier nommé bla et contenant des données au format txt.

1.1 Activité

Extensions courantes

Rendez-vous sur la page Wikipédia https://fr.wikipedia.org/wiki/Extension_de_nom_de_fichier et expliquer en quelques mots les extensions classiques suivantes :

- .jpg : .....
- .rar : .....
- .sh : .....
- .jpg : .....
- .mkv : .....
- .pdf : .....
- .ods : .....

2 Le stockage et la mise en forme

2.1 Les formats d'enregistrement

2.1.1 Le format .txt

un fichier texte est un fichier dont le contenu représente uniquement une suite de caractères ; il utilise nécessairement une forme particulière de codage des caractères qui peut être une variante ou une extension du standard local des États-Unis, l'ASCII. Il s'agit donc d'une suite de caractères, d'espace et de saut de ligne.
Aujourd'hui, la norme la plus utilisée pour les caractères est la norme utf8 : (abréviation de l'anglais Universal Character Set Transformation Format - 8 bits), Chaque caractère est repéré dans cet ensemble par un index entier aussi appelé « point de code », ce format permet ainsi 256 caractères. Il n'est pas suffisant pour coder tous les caractères spéciaux et tous les alphabets, c'est pourquoi il existe d'autres normes.

2.1.2 Le format .CSV

Le format csv est très courant sur internet et pour le stockage des données. Il est reconnu par de nombreux logiciels comme Excel, Libreoffice ainsi que n'importe quel éditeur de texte. Voici ce que nous dit Wikipédia sur le format CSV :

Comma-separated values, connu sous le sigle CSV, est un format informatique ouvert représentant des données tabulaires sous forme de valeurs séparées par des virgules.

Un fichier CSV est un fichier texte, par opposition aux formats dits « binaires ». Chaque ligne du texte correspond à une ligne du tableau et les virgules correspondent aux séparations entre les colonnes. Les portions de texte séparées par une virgule correspondent ainsi aux contenus des cellules du tableau.

⚠La virgule est un standard pour les données anglo-saxonnes, mais pas pour les données aux normes françaises. En effet, en français, la virgule est le séparateur des chiffres décimaux. Il serait impossible de différencier les virgules des décimaux et les virgules de séparation des informations. C'est pourquoi on utilise un autre séparateur : le point-virgule (;). À l'ouverture d'un fichier CSV il faudra souvent préciser le séparateur utilisé, si les données observées dans votre éditeur ne sont pas conformes à vos attentes, vérifiez le séparateur utilisé !

Les tableurs, tels que "Calc" (Libre Office), sont normalement capables de lire les fichiers au format CSV. J'ai précisé "normalement" car certains tableurs gèrent mal le séparateur CSV "point-virgule" et le séparateur des chiffres décimaux "virgule".

2.2 Les formats de mise en forme

À l'intérieur des fichiers les données sont souvent organisées sous forme de tableaux. On appelle **table** un tableau particulier dans lequel la première ligne sert à décrire le contenu des lignes suivantes.
Exemple : Le tableau suivant présente les couleurs associées aux différentes plages de longueurs d'onde d'une partie du spectre de la lumière.

La première ligne d'une table donne ainsi les **noms de champ ou critères** : le critère *longueur d'onde (nm)* indique que toutes les informations de la colonnes vont être des longueurs d'onde exprimées en nanomètre.

Longueur d'onde (nm)	Couleur
10-380	ultraviolet (invisible)
380-449	violet
449-490	bleu
490-573	vert
573-584	jaune
584-605	orange
605-700	rouge
700-3 000 000	infrarouge

TABLE 1 – Correspondance couleur-longueur d'onde

Avant les données brutes, le fichier commence usuellement par un *header* (une entête) qui contient des **métadonnées**.
Les **métadonnées** peuvent être enregistrées automatiquement lors de la génération du fichier comme par exemple, les métadonnées associées aux photos qui contiennent la date, l'heure, le lieu, les réglages de l'appareil photo lors de la prise de vue.

On appelle **collection de données ou série de données** l'ensemble des données qui ont le même critère.

★ Bilan ♡

Les **tables**, en informatique, servent à représenter des **collections de données** sous forme de colonnes (une série = une colonne). La première ligne permet d'explicité les **critères**.

2.3 Jouer avec les données

Il existe de nombreux logiciels permettant de modifier, d'indexer, de trier ou de visualiser les données autrement dit de **traiter les données**. Nous allons dans ce paragraphe nous focaliser sur les données contenues dans des tableaux. Les logiciels les plus courants pour réaliser ces actions sont Python, Libre Office Calc, Excell...

2.4 Trier

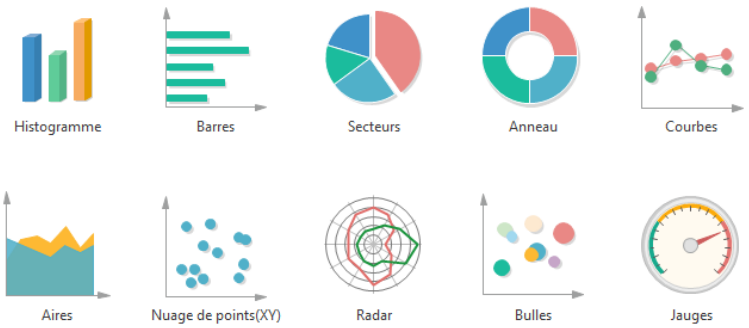
Trier, c'est réorganiser les lignes et les colonnes selon un critère précis. La modification d'une colonne implique souvent la modification de plusieurs autres colonnes pour que les données continuent à se correspondre.
Pour le cas du tableau de longueurs d'onde et de couleurs (table 1), on aurait pu choisir de classer les données par ordre décroissant de longueur d'onde : l'ordre des lignes des deux colonnes auraient alors été complètement inversé.

2.5 Filtrer

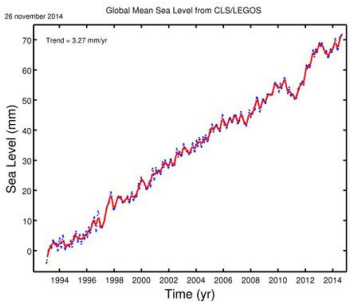
Filtrer, c'est sélectionner des données selon un critère particulier. Par exemple, si je veux conserver uniquement les informations concernant le spectre visible de la lumière, je vais donc filtrer (éliminer) toutes les longueurs d'onde supérieures à 700 nm et inférieures à 380 m. Je peux aussi **chercher** quelle longueur d'onde correspond à la couleur orange : la recherche est une opération de filtrage particulière.

2.6 Visualiser

L'une des façons les plus prisées pour représenter les données est la représentation graphique. Parmi les graphiques classiques on trouve :



Chaque type de graphique a ses particularités et ne convient pas à tous les types de données. C'est donc à l'utilisateur de trouver la forme la plus plus adaptée à ce qu'il veut mettre en valeur.
Par exemple : les histogrammes et secteurs vont permettre de représenter des pourcentages ou proportions alors que les évolutions temporelles seront plutôt représentée par des courbes.



★ Règle de mise en forme des graphiques ♡

Quelque soit la forme choisie les règles de mise en forme sont toujours les mêmes :

1. Le graphique doit être assez gros pour que tous les éléments soient lisibles.
2. Il doit être légendé : les noms des axes doivent apparaître (avec les unités utilisées) ainsi le sens des différentes couleurs doit être explicité.
3. Un titre expliquant l'intérêt du graphique doit être placé au dessus ou en dessous de la figure.

Application : A l'aide du tableur de votre choix, comparez la superficie des océans sachant que :
— l'océan Pacifique à une superficie de 165 250 000 km^2
— l'océan Atlantique, s'étend sur 106 400 000 km^2
— l'océan Indien couvre 73 556 000 km^2
— l'océan Arctique est le plus petit avec seulement 14 090 000 km^2
— l'océan Austral occupe 20 327 000 km^2
Réponse : Pacifique : 43.5% , Atlantique 28%, Indien 19.4%, Arctique 3.7%, Austral 5.4%, diagramme en secteur

3 Gérer des données avec Python

Voir activité Marée

4 Sauvegarder ses données

4.1 Cloud

Regardez la vidéo : https://www.youtube.com/watch?v=8y3psnPuH_o et répondre aux questions suivantes :

1. Qu'est ce que le Cloud ?
2. Quels types de service trouve-t'on sur le Cloud ?
3. Citer quelques fournisseurs Cloud.
4. Quel est l'intérêt de ce type de service ?
5. Où sont physiquement les données ?

4.2 Github et subversion

Aujourd'hui, la majorité des projets de développement libre et un grand nombre de projets dans les sociétés utilisent Subversion pour gérer leur code source. Ces plateformes vous permettent de choisir quand synchroniser vos données et quelles données synchroniser.
Ainsi, si vous réalisez un projet à plusieurs, vous aller créer sur votre ordinateur un dossier nommé projet que vous aller versionner (enregistré sur un serveur). Vous pouvez ensuite donner l'accès à ce dossier à d'autres utilisateurs qui pourront télécharger sur leurs ordinateurs le dossier et travailler dessus. Les modifications réalisées en *local* (sur les machines personnelles) ne sont envoyées au serveur que lorsque l'utilisateur réalise un *commit* (envoie). Les autres utilisateurs ne verront les modifications que s'ils exécutent la commande *update*.
L'intérêt majeur de ce système comparé au drive ou dropbox, c'est que chaque utilisateur est libre de modifier ce qu'il veut, sans impacter les autres jusqu'à ce qu'il considère que ses modifications sont suffisamment abouties pour être transmises. De plus, possède de nombreuses fonctionnalités comme la conservation des versions précédentes des documents !

4.3 Le coût écologique de nos données.

Commencez par regarder la vidéo suivante https://www.youtube.com/watch?time_continue=12&v=iiHxCX76bYU puis répondez aux questions.

1. Quelle quantité de données est stockée dans le serveur de l'entreprise citée ?
2. Qu'est-ce qu'un data center ?
3. Pourquoi la sauvegarde des données produit du CO_2 ?
4. Quelle est l'autre source de consommation d'énergie dans les datacenter ?
5. Comment sont gérées les pannes de courant électriques ?
6. Quelles sont les innovations "vertes" proposées par ce centre ? (en citer au moins 3).

Le coût d'une photo dans le cloud

Our recent piece on the carbon footprint of the internet generated plenty of coverage, so next up in our map of the world's carbon emissions is ... email. Of course, sending and receiving electronic message is never going to constitute the largest part of our carbon footprints. But the energy required to support our increasingly heaving and numerous inboxes does add up. Very roughly speaking (remember that all complex carbon footprints are really best guesses), a typical year of incoming mail for a business user – including sending, filtering and reading – creates a carbon footprint of around 135kg. That's over 1% of of a relatively green 10-tonne lifestyle and equivalent to driving 200 miles in an average car.
Extrait article The Guardian 21 octobre 2010

A long list of seemingly harmless everyday actions contribute to emissions of carbon dioxide (CO_2) and other climate-altering greenhouse gases. Driving a car and flipping a light switch have a clear "carbon footprint"—much less obvious is the harm caused by sending a simple text message or opening a bottle of water. Here is the environmental impact of some common activities :
Digital footprint :

Sending even a short email is estimated to add about four grammes (0.14 ounces) of CO2 equivalent (gCO2e) to the atmosphere. To put this into perspective, the carbon output of hitting "send" on 65 mails is on par with driving an average-sized car a kilometre (0.6 of a mile).
Extrait de Phyorg, November 26, 2015

Lire les extraits suivant et répondre aux questions :

1. De quand datent les deux articles ?
2. Repérer les mots qui vous semblent important et chercher leur sens en utilisant le dictionnaire en ligne *wordreference*.
3. Envoyer d'un email, est-ce un acte ecofriendly ?
4. Quelles sont les unités utilisées pour définir le coût énergétique ?
5. La situation semble t'elle avoir évoluée entre les deux articles ?
6. Faîtes vos propres recherche, quelle est la situation aujourd'hui ?

5 Conclusion

Rédiger un paragraphe de 20 lignes résumant ce que vous avez retenu de cette séquence en proposant une ouverture (débat) sur un sujet qui vous a marqué.