Thème 4 : Les données structurées

Introduction : Quelques repères historiques

Les données aussi ont une histoire !

– 1930 : utilisation des cartes perforées, premier support de stockage de données
– 1956 : invention du disque dur, qui permet de stocker de plus grandes quantités de données
– 2013 : Charte du G8 pour l'ouverture des données publiques
– 2016 : RGPD (Règlement Général sur la Protection des Données) qui protège les personnes dont on traite les données

1. La structuration des données

1.1 Données, descripteurs, métadonnées et collections

Ne confondez pas données et descripteurs !

- Prenons l'exemple d'un fichier de contacts : 07 99 99 99 99 peut être une **donnée** correspondant au numéro de téléphone d'un de vos amis.

« Numéro de téléphone » est le **descripteur** de cette donnée.

→ Donnez des exemples de descripteurs possibles qui permettent de caractériser un contact dans votre carnet d'adresse :

Adresse postale, adresse e-mail, date de naissance…

- Une **donnée** est donc la valeur prise par un **descripteur**, qui précise son sens. Un descripteur décrit donc le type de la donnée. Plusieurs descripteurs peuvent être utiles pour décrire un même objet.

- Une **métadonnée** est une donnée servant à définir ou décrire une autre donnée quel que soit son support (papier ou électronique). Par exemple, on associe à une donnée la date à laquelle elle a été produite ou enregistrée, ou à une photo les coordonnées GPS du lieu où elle a été prise.

Comment trouver des métadonnées ?

- **Sur une page du web**, pour accéder à la page source :

Sous Windows	Sous MacOS
1. Taper contrôle + U sur le clavier 2. Puis cherchez « meta »	1. Clic droit « Inspecter l'élément » 2. Puis cherchez « meta »

💡 Vous souvenez-vous que ces données se trouvent dans la balise <head> de votre fichier html ? Elles sont interprétées par votre navigateur, mais ne sont pas affichées dans la page, contrairement à ce qui se trouve dans la balise <body>…

- **Les métadonnées d'un fichier** stockés sur votre ordinateur sont accessibles en effectuant successivement les actions suivantes :

Sous Windows	Sous MacOS
1. Clic droit sur son icône 2. Choisir Propriétés 3. Cliquer sur l'onglet "Détails"	1. Clic droit sur son icône 2. Lire les informations

→ **Exercice** :

1. Lancez LibreOffice Writer (ou Google Docs) puis écrivez une phrase. Sauvegardez le fichier au format par défaut. Notez les métadonnées associées à ce fichier :

Type : Open document
Date : 2 avril 2024 à 13h20
Taille : 9 Ko

2. Lancez Microsoft Word (ou Apple Pages), refaites la même chose que précédemment. Comparez les métadonnées créées par les deux logiciels. Que remarquez-vous ?

Type : Document Word
Date : 2 avril 2024 à 13h23
Taille : 12 Ko

- Une **collection** regroupe des objets partageant les mêmes descripteurs, par exemple la collection des contacts d'un carnet d'adresse.

→ **Exercice :**

A partir d'un tableau, vous allez présenter une collection. Choisissez une collection comportant 8 objets :

1. Donnez un titre à votre collection.
2. Placez quatre objets en abscisse (prénoms de vos camarades par exemple…)
3. Notez deux descripteurs en ordonnée (sexe, taille, âge, au choix…)
4. Rentrez les données aux intersections.

Titre :

/	Arnaud	Bérénice	Chaïma	Driss
Sexe	M	F	F	M
Âge	14	15	15	16

1.2 Le Big Data

Le Big Data est un terme utilisé pour décrire l'abondance des données numériques et l'émergence des moyens développés pour y accéder et les analyser. Aujourd'hui le Big Data est déjà utilisé pour apprendre et résoudre des problèmes dans de nombreuses disciplines

comme l'intelligence artificielle. Les trois piliers du Big Data sont présentées à travers trois attributs surnommés les « 3 V » : le volume, la vélocité et la variété.

Le volume

Le **bit** (binary digit ou chiffre binaire) est la plus petite unité d'une donnée informatique. Un bit a une seule valeur binaire : 0 ou 1. Un **octet** (ou **byte** en anglais) est un « paquet » formé de 8 bits. Les quantités de mémoire sont exprimées en multiple d'octet.

1) Combien de nombres différents permet de coder un octet (c'est-à-dire 8 bits) ?

Le bit ne peut prendre que deux valeurs (0 et 1) et cela, huit fois dans un octet.
2 x 2 x 2 x 2 x 2 x 2 x 2 x 2 = 2^8 = 256 possibilités

2) Lisez le début de l'article https://fr.wikipedia.org/wiki/Préfixe_binaire qui vous explique pourquoi la valeur d'un Mo, d'un Ko a changé au cours du temps.

Avant 1998 :

Notation	Factorisation binaire	Nombre d'octets
1 octet (o)	1 o	1 octet
1 kilo-octet (Ko)	2^{10} octets	1024 octets
1 Méga-octet (Mo)	2^{20} octets	1 048 576 octets
1 gigaoctet (Go)	2^{30} octets	1 073 741 824 octets
1 téraoctet (To)	2^{40} octets	1 099 511 627 776 octets
1 pétaoctet (Po)	2^{50} octets	1 125 899 906 842 624 octets
1 exaoctet (Eo)	2^{60} octets	1 152 921 504 606 846 976 octets
1 zetaoctet (Zo)	2^{70} octets	1 180 591 620 717 411 303 424 octets
1 yotaoctet (Yo)	2^{80} octets	1 208 925 819 614 629 200 000 000 octets

Après 1998 :

Notation	Factorisation décimale	Nombre d'octets
1 octet (o)	1 o	1 octet
1 kilo-octet (Ko)	10^3 octets	1 000 octets
1 Méga-octet (Mo)	10^6 octets	1 000 000 octets
1 gigaoctet (Go)	10^9 octets	1 000 000 000 octets
1 téraoctet (To)	10^{12} octets	1 000 000 000 000 octets
1 pétaoctet (Po)	10^{15} octets	1 000 000 000 000 000 octets
1 exaoctet (Eo)	10^{18} octets	1 000 000 000 000 000 000 octets
1 zetaoctet (Zo)	10^{21} octets	1 000 000 000 000 000 000 000 octets
1 yotaoctet (Yo)	10^{24} octets	1 000 000 000 000 000 000 000 000 octets

On comprend bien que le second est plus pratique ! Le problème est que tous les systèmes n'utilisent pas le même : Windows utilise encore l'ancien alors que MacOS utilise le nouveau. Or, sur de très grandes valeurs, cela peut faire une différence allant jusqu'à 17% !

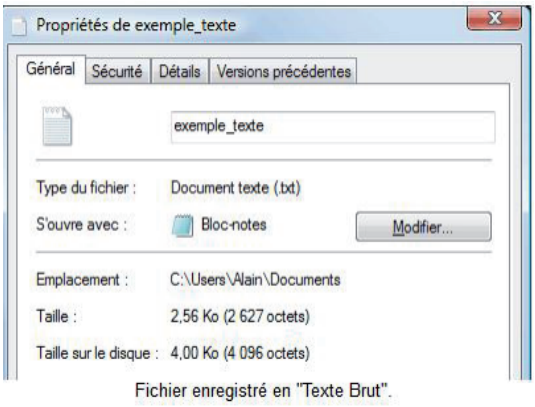
💡 Pour ne pas confondre les deux, on a renommé les anciennes valeurs (celles d'avant 1998) : kibioctet, mébioctet, gibioctet, tébioctet, etc.

3) Ranger dans l'ordre croissant ces quantités de données :

- Toutes les informations produites en 2013 : 5 Eo
- Un film de deux heures : 1 Go
- La NSA se dote pour 2013 d'un datacenter de 300 000 m^2 : 1 Yo
- 6 millions de livres : 1 To
- Volume de données mondiales en 2023 : 50 Zo
- Un morceau de musique au format MP3 : 5 Mo

5 Mo < 1Go < 1 To < 5 Eo < 50 Zo < 1 Yo

4) Voici les informations de métadonnées à propos d'un fichier :



a) Quel est le nom du document ?

exemple_texte

b) Quelle est son extension ?

.txt

c) Pour l'ordinateur, combien vaut un kilo-octet ?

La taille est de 4 Ko qui ici représentent 4096 octets. La taille est donc de 4096/ 4 = 1024.

d) Expliquer pourquoi (aidez-vous de l'exercice 2…)

Car sur Windows, l'ancienne notation a été conservée. Contrairement à MacOS, où 4 Ko = 4000 octets.

La vélocité

L'intérêt de la vitesse d'analyse des données est de pouvoir agir en temps réel. Pour chaque action en temps réel, définir quel type de données a été analysée.

1) Affichage d'une publicité personnalisée :

Des cookies

2) Équilibre de la température du thermostat :

Température issue d'un thermomètre/ capteur de température (sonde)

3) Surveiller l'intrusion dans un lieu :

Caméra de surveillance, capteur de mouvement, détecteur d'ouverture des portes…

La variété

Elle provient des différents types de données mais aussi de leurs différentes sources. Classer les éléments selon leur catégorie : source ou type de données (ex : un appareil photo numérique est une source, la photo qu'elle produit est un type de données).

image
montre
photo
numéro de téléphone
texte
thermostat
message vocal
smartphone
vidéo
article
tweet
navigateur GPS
document
coordonnées de géolocalisation
capteur

Source de données	***Type de données***
montre thermostat smartphone navigateur GPS capteur	image photo numéro de téléphone texte message vocal vidéo article tweet document coordonnées de géolocalisation

→ Nous voyons que les sources de données sont des objets et les types de données sont des informations qui peuvent être dématérialisées.