Séance 03 : Les données structurées et leur traitement

La séance repose sur l'utilisation du logiciel MP3tag téléchargeable ici :
https://www.mp3tag.de/en/download.html
Ce logiciel permet de modifier les tags (métadonnées) des fichiers musicaux à partir de bases de données en ligne et /ou de les exporter au format csv.
Il s'agira d'explorer et de modifier les métadonnées de fichiers de musique.

Activité d'introduction : Comment documenter une playlist ?

On dispose, dans un répertoire, de musiques libres de droit téléchargées au préalable sur le site https://www.auboutdufil.com/

Sans ouvrir aucun de ces fichiers, quelles informations pouvez-vous en tirer ?

Le format **mp3** laisse à penser qu'ils contiennent de la musique.
Il s'agira d'ouvrir le dossier « musiques libres de droit » et de cliquer droit sur chaque fichier de musique (mp3).

Un **« clic droit propriété »** donne accès aux métadonnées

D'ailleurs un « clic droit » sur le bandeau du dossier donne accès à toutes les métadonnées possibles à renseigner sous windows (sélectionnées par défaut en fonction du format)

Ouvrir le logiciel MP3tag, et à partir du logiciel, ouvrir le dossier « Séance 03 - Fichiers musicaux bruts »

Les fichiers musicaux avec quelques métadonnées apparaissent.

Faire une recherche internet pour compléter à la main les tags (**métadonnées**) du fichier « Bill_CheathumShake that Little Foot_64kb » à partir des infos trouvées, dans l'espace de saisie.

A l'aide du logiciel, faire une recherche dans la base de données « MusicBrainz » afin de compléter les tags (métadonnées) des autres musiques.

Exporter les métadonnées au format CSV de façon à avoir une vision globale de la médiathèque dans un tableur.

Comment faire avec un très grand nombre de données ?

Enregistrer le fichier csv au format texte et le placer dans le même répertoire que lecture métadonnées.py

Veillez à la concordance des noms de fichiers appelés par le programme, lisez-le avec Python.

Souvent les bases sont tellement grandes que le tableur ne suffit pas à les traiter. Le format texte ou csv est peu gourmand en taille et peut être ouvert à l'aide d'un programme.

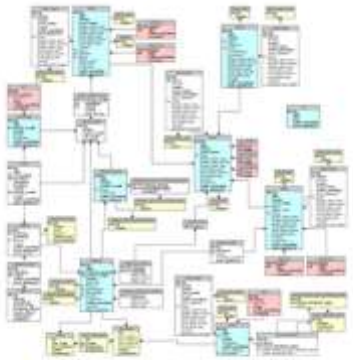
Etude de la base de données Music brainz

Aller sur le site https://musicbrainz.org/

Trouver la licence sous laquelle les données sont publiées et expliciter ce que l'on peut en faire.

Trouver les descripteurs principaux utilisés ainsi que les descripteurs secondaires. Trouver le nombre d'artistes référencés.

Le schéma des principaux descripteurs illustre la complexité d'une telle base et justifie l'utilisation d'un programme.



https://musicbrainz.org/doc/MusicBrainz_Database/Schema

Un autre enjeu de telles bases de données : l'intelligence artificielle

Vous avez:

- L'historique d'écoute complet pour 1 million d'utilisateurs
- La moitié de l'historique d'écoute pour 110 000 autres utilisateurs
- Vous devez prédire la moitié manquante.

The Million Song Dataset Challenge : https://www.kaggle.com/c/msdchallenge

Il s'agit de créer un algorithme qui pour un utilisateur donné va étudier son historique d'écoute. A partir de cette étude pour un million d'utilisateurs (set d'apprentissage), il s'agira d'établir un lien entre les musiques écoutées précédemment et les suivantes.
Une fois ce lien établi, il s'agira à partir de la moitié de l'historique d'écoute de 110 000 utilisateurs de reconstruire à l'aide de l'algorithme établi précédemment l'autre moitié. Les prédictions les plus proches de l'historique réel gardé secret gagnent.