Dépôt 1 : Nettoyage des données

Nettoyer la colonne *ville* en enlevant les informations du style “(y)” [ex. (fleuve)] et aussi par exemple “ -- château ” puis garder seulement les villes (ex. pas de fleuves (Loire), etc.) ~~et enlever les doublons~~.

Maintenant on peut créer des nouvelles lignes à partir des villes (s’il y a plus que 1 ville) pour que la colonne *ville* n’est plus décomposable (ex. “Chécy, Viglain, Tigy” → créer 3 lignes avec les mêmes informations/contenus des colonnes).

Création des nouvelles lignes en séparant les références du style “ xx | xx.
Les lignes vont être créés par rapport aux séparations “ | “ dans les données et s’il n’y en a pas, les informations vont être simplement doublé (→ référence peut être integer maintenant).

Suppression des informations de pouces dans la colonne *tailles_du_cliché* (on pourra les calculer à partir de la longueur et largeur [expliqué plus bas]).

Création des nouvelles lignes en séparant les *nombre_de_cliché* avec leur *taille_du_cliché* respectif (ex. “1, 2” ; “6x7, 24x36” → cela va donner deux lignes). Séparation en lien avec les colonnes *négatif_ou_inversible* et *couleur_ou_noir_et_blanc* qui vont être séparés par rapport à la virgule.

Suppression du texte “Prise de vue:” dans la colonne *date*.
Transformer janvier en 01, février en 02, etc.
Puis création de nouvelles lignes à partir d’où il y a plusieurs dates pour une seule référence (ex.: “octobre 1983/janvier 1984” → copier l’ensemble des informations, ex.2: “octobre-novembre 1972”) → rendre le format date possible.

~~Suppression des doublons dans les colonnes *sujet*, *index_iconographique*, *index_personnes*~~. Création de nouvelles lignes également (plus optimisé).

Enlever tous les informations inutiles qui ne contiennent pas “négatif” ou “inversible” dans la colonne *négatif_ou_inversible* (pareil pour la colonne *couleur_ou_noir_et_blanc*).

Enlever tous les informations dans *index_iconographique* qui sont déjà présentes dans la colonne *sujet* (en raison de thesaurus c’est dans ce sens).

Enlever tous les informations dans *notes_de_bas_de_page* qui sont déjà présents dans les autres colonnes, et la même chose pour la colonne *index_personnes* (informations déjà présentes dans *sujet* et *index_iconographique*).
Enlever les champs qui ont seulement des virgules dans la colonne *sujet* et les mettre à null et enlever les virgules dans les champs qui commence par une virgule.

Réunir les lignes qui sont exactement pareils et faire une somme des nombre de clichés.

Vider la colonne *discriminant* car elle n'est pas très utile pour l'instant avec seulement ~ 20 entrées donc de suite pour chaque numéro d'article qui apparaît une deuxième, troisième, …, n fois (boucle pour trouver), on va mettre un discriminant (a,b,c,...) pour pouvoir créer une clé primaire (article, discriminant).

Création d'une colonne *id_ville* pour pouvoir stocker les noms des villes dans une table à part, avec un *identifiant* qui va pouvoir être référencé (gain de lisibilité et d'espace mémoire). Ajout des colonnes *code_postal*, *long*, *lat* (lambert 93) pour la table des villes.

Création d'une colonne *id_taille_du_cliché* pour faire une table qui contient un *id*, la taille séparé en *longueur* et *largeur*. On va pouvoir faire cette table à cause des tailles déjà prédéfinis (Formats de plans-films).

Création d'une colonne *id_serie* pour faire une table qui contient un *id* et le *nom* de la série (gain d'espace mémoire et on ne pourra pas se tromper dans l'écriture).