Point de contrôle qualité	Problème observé et commentaire	Correction : comment ? importance ?
1. le jeu de données est difficilement accessible (format “image”, PDF, HTML...), le fichier est mal formé	Non : format CSV disponible Possibilité de l’exporter dans d’autres formats : Json, Excel, GeoJSON, Shapefile, KML.	
2. la licence est absente ou inhabituelle, le jeu de données n’est pas “open data”	Non : Licence Ouverte v2.0 (Etalab)	
3. le fichier fait peu appel aux standards répandu et aux données pivots	Non	
4. le fichier est mal documenté	Peu de documentation sur le jeu de données sur la plateforme data.culture.gouv.fr directement. Mais présence de liens renvoyant vers différentes documentations sur le projet Joconde.	
5. il existe des problèmes de syntaxe	Choix de ne pas contrôler le fichier de départ de plus de 600 000 lignes, sauf pour les colonnes stratégiques (dates, localisation).	Choix de faire un contrôle qualité sur les partitions du corpus pour plus de lisibilité.

Point de contrôle qualité	Problème observé et commentaire	Correction : comment ? importance ?
6. Valeurs aberrantes, suspectes, inexplicables, pas crédibles	Choix de ne pas contrôler le fichier de départ de plus de 600 000 lignes, sauf pour les colonnes stratégiques (dates, localisation).	Choix de faire un contrôle qualité sur les partitions du corpus pour plus de lisibilité.
7. Il manque des données et cela n'est pas documenté (trous, données tronquées, valeurs vides, granularité / fréquence / maillage / fraîcheur)	Un grand nombre de colonnes, qui sont inégalement saisies selon les notices, les musées et les types d'œuvres. De plus : types de colonnes ne sont pas forcément applicables à tous les types d'œuvres. Un nombre très conséquent de valeurs vides, ce qui constitue la principale difficulté lors de l'exploitation de ce jeu de données.	Impacte la possibilité de faire des visualisations, car seulement un petit nombre de colonnes sont saisies de façon régulière. Seules les colonnes relatives à la localisation géographique du musée sont complètes. Solution : faire des partitions du corpus initial pour isoler des corpus de taille plus réduite, permettant de traiter au mieux la question des valeurs vides.
8. Trop de données : doublons, inutilement vieilles, précision / fréquence / maillage / fraîcheur	Plusieurs colonnes relatives à la datation : Periode_de_creation, Millesime_de_creation, Epoque. Donc une information sur la datation de l'œuvre saisie dans 3 colonnes différentes, et dans des formats de date différents. Fréquence de mise à jour hebdomadaire tous les mercredis.	Solution pour les dates: retraitement des colonnes de datation pour harmoniser le format de date.
9. Données posant problème avec la réglementation (données perso, relatives à la santé, la religion..., propriété littéraire et artistique, etc.)	Non	
10. Les contenus posent problèmes : synonymies, non traduits (USA), cryptique (DAECPP), utilisation du 0 au lieu du "null"...	Non	