| ID | N° d'ordre | Domaine du problème | Types de données concernées | Type de problème | Exemple | Détecter le problème | Ligne de commande pour le contrôle | Réduit considérablement l'usage des données | Niveau d'importance du problème | Temps d'analyse en min | Analyse semi-automatisable | Solution | Temps de traitement | Remarques |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | Jeu de données & métadonnées | Toutes | Le jeu de données est dans un format "image" ne permettant pas de manipuler les données | Le jeu de données est un fichier image au format JPEG ou PDF. | Ouvrir le fichier et tenter de copier/coller les données | | oui | 1 | 1 | 1 | * Demander au producteur une version qui permette de manipuler les données (CSV, Excel, etc.)<br>* Essayer une phase d'OCR du document | | |
| 2 | 2 | Jeu de données & métadonnées | Toutes | Le jeu de données est dans un format non spécifiquement adapté aux données : PDF, Word, ODF, epub, HTML, SVG, etc. | Le jeu de données est un fichier HTML. | Déterminer le format du fichier | | oui | 1 | 1 | 1 | Dans certains cas la méthode du scrapping est une solution. | | Les formats PDF ou de traitement de texte rendent l'exploitation des données difficiles. |
| 3 | 3 | Jeu de données & métadonnées | Toutes | Le format du jeu de données n'est pas précisé (fichier CSV, TSV, etc.) | l'extension du jeu de données ne permet pas de savoir quel logiciel permet de l'ouvrir et l'éditeur n'a pas fourni d'indication complémentaire | Essayer d'ouvrir le fichier ? | | | 1 | 1 | 0 | * Demander au producteur<br>* Rétro-documenter le format | | |
| 31 | 4 | Jeu de données & métadonnées | Toutes | La documentation et les métadonnées sont quasi inexistantes voire absentes | La documentation tient sur 5 lignes alors que le fichier est très complexe | Lecture des métadonnées | | oui | 1 | 1 | 0 | * Produire les métadonnées<br>* Faire renseigner ou valider les métadonnées par le producteur | | C'est important de voir ça très en amont : si l'on veut corriger cela, certains points de contrôle participeront à la documentation |
| 115 | 5 | Jeu de données & métadonnées | Toutes | La documentation et les métadonnées sont d'un usage difficile (doc papier, doc au format PDF image, doc uniquement en anglais, etc.) | La documentation est fourni sous forme de PDF image : les usagers ne peuvent pas rechercher des termes pour y naviguer rapidement | Consultation de la doc et des métadonnées | | | 1 | 1 | 0 | * Demander au producteur | | |
| 4 | 6 | Jeu de données & métadonnées | Toutes | La licence du jeu de données ne nous permet pas de l'utiliser | Le jeu de données est un fichier commercial que l'on n'a pas acheté | En cas de doute, demander au producteur d'où viennent les données | | | 1 | 15 | 0 | | | |
| 5 | 7 | Jeu de données & métadonnées | Toutes | Le format du jeu de données n'est pas ouvert | Le fichier est au format .xls ou .xlsx | Ne pas seulement se baser sur l'extension mais ouvrir également le fichier | file nom.du.fichier | | 1 | 1 | 1 | * Vérifier que le document existe dans un format ouvert<br>* Convertir le document dans un format ouvert | | |
| 6 | 8 | Jeu de données & métadonnées | Toutes | Le format du jeu de données ne permet pas d'ouvrir le fichier dans des outils très répandus (Excel, Notepad...) | Le fichier au format .csv s'ouvre mal dans Excel, outil le plus répandu pour ouvrir des tableaux | Essayer d'ouvrir le fichier ? | | | 1 | 1 | 1 | | | |
| 7 | 9 | Jeu de données & métadonnées | Toutes | L'encodage du fichier n'est pas spécifié (ISO-8859-1, UTF8, etc.) | Le fichier contient des caractères ésotériques mais on ne sait pas s'il s'agit d'un problème d'encodage | Lecture des métadonnées | | | | 1 | 0 | * Produire les métadonnées<br>* Faire renseigner ou valider les métadonnées par le producteur | | |
| 8 | 10 | Jeu de données & métadonnées | Toutes | L'encodage n'est pas en UTF-8 : ce dernier devient la norme de facto et d'autres encodages peuvent engendrer des problèmes | L'encodage est en ISO-8859-1 | Ouvrir le fichier avec un éditeur qui spécifie l'encodage | file file.csv | | | | | iconv -f ISO-8859-1 -t UTF-8 . /caracteristiques_2015.csv > caracteristiques_2015_rev1.csv | | |
| 9 | 11 | Jeu de données & métadonnées | Toutes | L'encodage n'est pas homogène | Certaines données sont correctement encodées et d'autres contiennent des caractères ésotériques | Ouvrir le fichier et parcourir visuellement les données ; rechercher quelques chaînes comme "Ã©" ou "?" ou "?t?" ou "g? n?ral" etc. | perl -F/,/ -alne 'print if 1..1;print if m/([a-z]{3}[a-z]Ã©)\;' file.csv \| csvsort \| csvlook | | | | | | | |
| 10 | 12 | Jeu de données & métadonnées | Toutes | Le fichier est mal formé | Pour certaines lignes, parfois une colonne manque, ou, le fichier cvs comporte des "virgule" non formatées et empêche l'ouverture correcte du fichier | Ouvrir le fichier, trier la dernière colonne du fichier et regarder le résultat | csvclean --dry-run file.csv | oui | 1 | 1 | 1 | | | |
| 11 | 13 | Jeu de données & métadonnées | Toutes | Le jeu de données concernant des horaires de mode de transport ne possède pas de version au format GTFS | Le fichier n'est pas au format GTFS | | | | 2 | 1 | 1 | | | |
| 12 | 14 | Jeu de données & métadonnées | Toutes | Le jeu de données concernant des œuvres n'est pas au format Dublin Core | Le fichier n'est pas au format Dublin Core | | | | 2 | 1 | 1 | | | |
| 13 | 15 | Jeu de données & métadonnées | Toutes | Le jeu de données utilise une norme peu accessible au plus grand nombre (coût, complexité) | Le jeu de données est au format TRIDENT | | | | 2 | 1 | 0 | | | |
| 14 | 16 | Jeu de données & métadonnées | Toutes | Le process d'acquisition n'est pas connu | Wiki leaks | Lecture des métadonnées | | | 2 | 1 | 0 | Évaluer le process et vérifier un échantillon de données | | Redondant avec "Métadonnées imprécises : process et contexte de production non explicités" ? |
| 15 | 17 | Jeu de données & métadonnées | Toutes | L'échantillon n'est pas documenté | L'échantillon semble représentatif mais on ne peut pas vérifier qu'il le soit bien, puisque ce dernier n'est pas documenté | Lecture des métadonnées | | | 2 | 1 | 0 | | | |
| 16 | 18 | Jeu de données & métadonnées | Toutes | Le format d'un des champs n'est pas documenté, si bien qu'on ne peut comprendre ce qu'il contient ou bien contrôler ses valeurs | * La date est parfois exprimée par le nombre de secondes depuis 1970 ; cette donnée est difficile à comprendre.<br>* Un jeu de données contient un champ "Image" en binaire, dont le format n'est pas spécifié. | Lecture des métadonnées et ouverture du fichier : le format champ binaire est-il documenté ? | | | 2 | 3 | 0 | * Demander au producteur<br>* Rétro-documenter le format | | |
| 17 | 19 | Jeu de données & métadonnées | Toutes | La taille maximale d'un champ n'est pas documentée | On ne sait pas si un code peu dépasser 10 caractères et si certaines valeurs sont donc fausses | Lecture des métadonnées | | | 2 | 3 | 0 | * Demander au producteur<br>* Rétro-documenter le format | | |
| 18 | 20 | Jeu de données & métadonnées | Toutes | Pour tel champ, l'incertitude de la mesure n'est pas documentée (appelée aussi "précision", exprimée en % ou bien "à plus plus ou moins X unités près") | Des coordonnées GPS sont indiquées mais on ne connaît pas leur marge d'erreur (précises à 10 m, à 100 m ?) ; la précision d'une mesure de température n'est pas explicitée (+/- 0,1° ? +/- 1° ?) | Lecture des métadonnées | | | 3 | 1 | 0 | | | |
| 117 | 21 | Jeu de données & métadonnées | Toutes | L'incertitude de la mesure n'est pas connue par le producteur | Le producteur des données ne connaît pas la précision de ses mesures | Si l'incertitude de la mesure n'est pas documentée (ID18), demander au producteur | | | 3 | 15 | 0 | | | |
| 22 | 22 | Jeu de données & métadonnées | Toutes | La précision n'est pas cohérente avec la granularité : l'incertitude de la mesure est 100 fois supérieure à la granularité | Des coordonnées géographiques annoncent une granularité au cm alors que l'incertitude des appareils de mesure est de +/- 5 mètres | Lecture des métadonnées | | | 3 | | 0 | | | |
| 19 | 23 | Jeu de données & métadonnées | Toutes | L'origine de certaines données est une entrée manuelle non contrôlée | Le risque est d'obtenir 25 orthographes de "Saint-André-des-Arts" | Ouvrir le jeu de données et parcourir : des données sont-elles manifestement entrées à la main ? | | | 3 | 3 | 0 | | | |
| 20 | 24 | Jeu de données & métadonnées | Toutes | Les données proviennent d'un processus de reconnaissance automatique dont la marge d'erreur est globalement bonne mais localement problématique (OCR, reconnaissance de forme, géocodage, etc.) | OCR ;<br>reconnaissance automatique des visages (va dépendre de la qualité de la lumière de la prise de vue, de la couleur des personnes concernées (c'est encore un problème en 2016)) ; etc. | Ouvrir le jeu de données et parcourir : des données sont-elles manifestement issues d'un processus de reconnaissance automatique ? | | | 3 | 3 | | | | |
| 21 | 25 | Jeu de données & métadonnées | Toutes | L'échantillon est biaisé | Certaines populations sont absentes, sur-représentées ou sous-représentées ; les données subissent une forte variation saisonière | WTFcsv | | | 3 | | 0 | | | |
| 23 | 26 | Jeu de données & métadonnées | Toutes | Le process de signalement d'erreur et d'échange avec le producteur n'est pas explicité | Aucune forme de contact n'est donnée | Lecture des métadonnées | | | 2 | 1 | 0 | * Produire les métadonnées<br>* Faire renseigner ou valider les métadonnées par le producteur | | |
| 24 | 27 | Jeu de données & métadonnées | Toutes | Le process de signalement d'erreur et d'échange avec le producteur n'existe pas ou bien il est défaillant | Le producteur ne répond pas aux questions | Demander au producteur | | | | 15 | 0 | | | |
| 25 | 28 | Jeu de données & métadonnées | Toutes | La disponibilité de la donnée n'est pas documentée (temps pendant lequel la donnée est accessible par rapport au temps total souhaité, généralement exprimé en pourcentage) | L'utilisateur ne sait pas si la qualité de service est de 95% ou 99,99%. Si le système qui héberge la donnée est régulièrement inacessible (maintenance, etc.), les usagers devraient en être informé pour savoir si leur usage en est impacté | Lecture des métadonnées | | | | 1 | | | | https://fr.wikipedia.org/wiki/Disponibilit%C3%A9 |
| 26 | 29 | Jeu de données & métadonnées | Toutes | La disponibilité de la donnée n'est pas mesurée | Le producteur ne sait pas si la qualité de service est de 95% ou 99,99 alors que tel futur usage est critique | Demander au producteur | | | | 15 | 0 | | | https://fr.wikipedia.org/wiki/Disponibilit%C3%A9 |
| 116 | 30 | Jeu de données & métadonnées | Toutes<br><br>Le mode d'accès à la donnée est un frein à l'usage (temps d'accès, droit d'accès long et complexe, droit d'accès limité) | La requête d'une donnée "temps réel" met plus de 40 secondes ; l'accès à la donnée nécessite un certificat de sécurité long à obtenir ; l'architecture du site ne permet pas à un robot de télécharger les actualisations des données | * Tester l'accès aux données<br>* Tester la récupération des données via un outil automatisable (commande wget par exemple) | | | 2 | 5 | 1 | | | | |
| 27 | 31 | Jeu de données & métadonnées | Toutes | La mesure de la qualité n'est pas documentée | Des contrôles qualité existent (amont ou aval) mais ils ne sont pas explicités si bien qu'on ne peut savoir si tel champ est fiable ou non | Lecture des métadonnées | | | | | | | | |
| 28 | 32 | Jeu de données & métadonnées | Toutes | La qualité de la donnée n'est pas mesurable à travers des contrôles formels | Il n'existe pas de méthode de contrôle permettant de dire si la syntaxe de ce champ est bonne | Demander au producteur | | | | | 0 | | | |
| 29 | 33 | Jeu de données & métadonnées | Toutes | La qualité de la données n'est pas mesurée | Aucune méthode de contrôle n'est mise en oeuvre pour mesurer la qualité des données | Demander au producteur ou à l'éditeur | | | 3 | 15 | 0 | | | |
| 30 | 34 | Jeu de données & métadonnées | Toutes | Une entité possède plusieurs identifiants | | | | | | | 0 | | | Exemple que me prend Simon sur les Asso qui ont à la fois un numéro d'asso et un code SIREN. |
| 32 | 35 | Jeu de données & métadonnées | Toutes | Le nom ou titre du jeu de données est vague, ambigu ou trop complexe : titre de la notice éditoriale, nom donné dans les métadonnées ou dans la documentation (pas le nom du fichier) | * "Résultat des élections" : lesquelles ? où ? quand ?<br>* "Résultats des élections à Montréal" : il existe 6 communes appelées Montréal dans le monde... | Lecture des métadonnées, de la documentation et/ou de la fiche de présentation | | oui | 1 | 1 | 0 | | | |
| 33 | 36 | Jeu de données & métadonnées | Date | Manque de métadonnées : fourchette temporelle non explicitée | Des dates figurent dans le jeu mais aucune métadonnée ne peut confirmer la fourchette attendue de ces dates. Exemple : Trésorerie du 01/02/2010 au 24/11/2016. | Lecture des métadonnées | | | 2 | 1 | 0 | * Produire les métadonnées<br>* Faire renseigner ou valider les métadonnées par le producteur | | |
| 34 | 37 | Jeu de données & métadonnées | Date | Manque de métadonnées : zone spatiale non explicitée | Des coordonnées figurent dans le jeu mais aucune métadonnée ne peut confirmer la zone d'appartenance attendue pour ces points. | Lecture des métadonnées | | | 2 | 1 | 0 | * Produire les métadonnées<br>* Faire renseigner ou valider les métadonnées par le producteur | | |
| 35 | 38 | Jeu de données & métadonnées | Nombre | Manque de métadonnées : fourchette non spécifiée | On peut attendre d'un nombre qu'il soit compris entre une valeur minimum et une valeur maximum ; par exemple l'âge d'une personne devrait toujours être entre 0 et 130 voir 18 et 70 selon les cas. | Lecture des métadonnées | | | 2 | 1 | 0 | * Produire les métadonnées<br>* Faire renseigner ou valider les métadonnées par le producteur | | |
| 36 | 39 | Jeu de données & métadonnées | Booléen | Manque de métadonnées : le fait que le champ soit un booléen n'est pas spécifié | | Lecture des métadonnées | | | 2 | 1 | 0 | * Produire les métadonnées<br>* Faire renseigner ou valider les métadonnées par le producteur | | |
| 37 | 40 | Jeu de données & métadonnées | Booléen | Manque de métadonnées : le format du booléen n'est pas spécifié | On ne sait pas à quelles valeurs s'attendre : "vrai"-"faux" ou "oui"-"non" | Lecture des métadonnées | | | 2 | 1 | 0 | * Produire les métadonnées<br>* Faire renseigner ou valider les métadonnées par le producteur | | |
| 38 | 41 | Jeu de données & métadonnées | Toutes | Manque de métadonnées : processus et contexte de production non explicités | On ne sait pas si une mesure vient d'un capteur ou d'une mesure manuelle | Lecture des métadonnées | | oui | | 1 | 0 | * Produire les métadonnées<br>* Faire renseigner ou valider les métadonnées par le producteur | | |
| 39 | 42 | Jeu de données & métadonnées | Toutes | Manque de métadonnées : la fraicheur des données n'est pas explicité :<br>* le délai entre le réel et la mise en base de la donnée<br>* le délai entre le réel et la publication de la donnée | Il n'est pas dit si telle information sur une grossesse va mettre plus de neuf avant d'arriver au réutilisateur | Lecture des métadonnées | | | | 1 | 0 | * Produire les métadonnées<br>* Faire renseigner ou valider les métadonnées par le producteur | | |
| 40 | 43 | Jeu de données & métadonnées | Toutes | Manque de métadonnées : la langue des textes n'est pas spécifiée | | Lecture des métadonnées | | | | 1 | 0 | * Produire les métadonnées<br>* Faire renseigner ou valider les métadonnées par le producteur | | |
| 41 | 44 | Jeu de données & métadonnées | Date | Métadonnées imprécises : le format de date n'est pas spécifié | Format américain ? anglais ? européen ? etc. | Lecture des métadonnées | | | | 1 | 0 | * Produire les métadonnées<br>* Faire renseigner ou valider les métadonnées par le producteur | | |
| 42 | 45 | Jeu de données & métadonnées | Nombre | Métadonnées imprécises : unités non spécifiées | On ne dit pas si colonne "hauteur" est en cm ou dm | Lecture des métadonnées | | oui | 1 | 1 | 0 | * Produire les métadonnées<br>* Faire renseigner ou valider les métadonnées par le producteur | | |
| 43 | 46 | Jeu de données & métadonnées | Coordonnées | Métadonnées imprécises : système de coordonnées non spécifié | La documentation n'indique pas si les coordonnées sont en WGS 84, Lambert ou un autre système | Lecture des métadonnées | | | | 1 | 0 | * Produire les métadonnées<br>* Faire renseigner ou valider les métadonnées par le producteur | | Redondant avec "Document : Le process d'acquisition n'est pas connu" |
| 44 | 47 | Jeu de données & métadonnées | Toutes | Métadonnées imprécises : nom de colonnes ambigus | "Emplacement" ne dit rien sur la donnée attendue : une adresse ? "en haut" ? "devant" ? etc. | Lecture des métadonnées | | oui | | 1 | | * Produire les métadonnées<br>* Faire renseigner ou valider les métadonnées par le producteur | | |
| 45 | 48 | Jeu de données & métadonnées | Toutes | Métadonnées fausses | | Lecture des métadonnées | | | | | | | | |
| 46 | 49 | Jeu de données & métadonnées | Toutes | La taille maximale d'un champ dépasse celle qui est spécifiée dans la documentation | La colonne "âge" spécifie une longueur de 3 caractères maximum et certaines valeurs sont des 4 caractères ou plus | Lecture des métadonnées et des données | csvgrep -c colonne_x -r ".{25,}" file.csv \| csvlook<br>csvsql --query "SELECT MAX (LENGTH(mois)) FROM file" file.csv | | 3 | 3 | 1 | | | |
| 47 | 50 | Jeu de données & métadonnées | Toutes | L'ordre des colonnes ne correspond pas à l'ordre donné dans la documentation | La documentation donne Prénom;Nom;Âge; Profession alors que le jeu se présente sous la forme Nom;Prénom;Age;Profession | Lecture des métadonnées et des données | | | 3 | 3 | 0 | | 5 | |
| 48 | 51 | Jeu de données & métadonnées | Coordonnées | Les coordonnées ne sont pas au format WGS 84 | Les coordonnées sont au format Lambert II nécessitant une conversion des points pour des usages mobile liés à des GPS grand public | Lecture des métadonnées et des données | | | 3 | 3 | | | | |
| 49 | 52 | Jeu de données & métadonnées | Chaîne alpha | Les codes pays ne sont pas au format ISO 3166 alpha-3 | L'Allemagne est noté "ALL" alors qu'il existe un code ISO employé internationalement | Lecture des métadonnées et des données | | | 2 | 3 | 1 | | | |
| 50 | 53 | Jeu de données & métadonnées | Chaîne alpha | Les codes de langues ne sont pas au format ISO 639 | Le français est noté "F" ou "français" en lieu et place de "fr" | Lecture des métadonnées et des données | | | 2 | 3 | 1 | | | |
| 51 | 54 | Jeu de données & métadonnées | Date | Les dates ne sont pas au format ISO 8601 | La date est notée "01/01/2016" | Lecture des métadonnées et des données | | | 2 | 3 | 1 | | | |
| 52 | 55 | Jeu de données & métadonnées | Chaîne alpha<br><br>Les monnaies ne sont pas au format ISO 4217 | Le franc suisse est noté FS | Lecture des métadonnées et des données | | | | 3 | 3 | 1 | | | https://fr.wikipedia.org/wiki/ISO_4217 |
| 53 | 56 | Syntaxe | Chaîne alpha | Erreurs syntaxiques : espace(s) au début ou à la fin du champ | " Pierre" au lieu de "Pierre" | Rechercher les cas à partir d'une regexp | perl -F/,/ -alne 'print if 1..1;print if m^ .*[\^ ]*$/\;' file.csv \| csvsort \| csvlook | | | 3 | 1 | | | |
| 54 | 57 | Syntaxe | Chaîne alpha | Erreurs syntaxiques : bug syntaxique dans les strates du SI : le cas de l'apostrophe | "N'Diaye" à la place de N'Diaye | Rechercher "\'" | csvgrep -c colonne_x,colonne_y -m \' file.csv | | | 1 | 1 | Remplacer \' par ' | | |
| 55 | 58 | Syntaxe | Nombre | Erreurs syntaxiques : syntaxe des numéros ou nombres en tous genres | "45€" au lieu de "45"; "1,000,000" au lieu de "1000000" | | | | | 30 | | | | |

| ID | N° d'ordre | Domaine du problème | Types de données concernées | Type de problème | Exemple | Détecter le problème | Ligne de commande pour le contrôle | Réduit considérablement l'usage des données | Niveau d'importance du problème | Temps d'analyse en min | Analyse semi-automatisable | Solution | Temps de traitement | Remarques |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 56 | 59 | Syntaxe | Chaîne alpha | Erreurs syntaxiques : codes (code INSEE, code postal, SIRET, SIREN, n° de Sécu, ISBN, ISSN, IBAN, BIC, code ROM, indicatif du pays, code APE, code NAF, etc.) | 7100 au lieu de 07100 pour un code postal | Bâtir la regexp relative au code attendu et tester. | csvgrep -c colonne_x -r "(^(\d\d\d\d\d)" file.csv \| csvsort \| csvlook # code postal | | | 5 | 1 | | | |
| 57 | 60 | Syntaxe | Chaîne alpha | Erreurs syntaxiques : sigles et abréviations | "SNCF", "S. N. C. F.", "S.C.N.F." ? "Bouf" ou "Boul." ou "Bid" ? | Classer la colonne date par ordre alphabétique permet de rapidement voir les problèmes de syntaxe | | | | 5 | | | | |
| 58 | 61 | Syntaxe | Booléen | Erreurs syntaxiques : booléen | "V" au lieu de "1" selon la spécification du booléen | Regarder WTFcsv sur les colonnes de booléens | csvstat -c col_booleen --freq file.csv | | | 1 | 1 | | | |
| 59 | 62 | Syntaxe | Chaîne alpha | Erreurs syntaxiques : email, url | laurent.dupont@wanadoo@fr | | | | | | 1 | | | |
| 60 | 63 | Syntaxe | Date | Erreur syntaxique sur la date | 2016/09/30 au lieu de 2016-09-30 attendu | Classer la colonne date par ordre alphabétique permet de rapidement voir les problèmes de syntaxe | csvcut -c colonne_x file.csv \| csvsort \| csvlook | | | 1 | 1 | | | |
| 61 | 64 | Syntaxe | Chaîne alpha | Incohérences syntaxiques : syntaxe des noms propres | "de La Tour" ou "La Tour (de)" ? | | | | | 5 | 1 | | | |
| 62 | 65 | Syntaxe | Nombre | Incohérences syntaxiques : homogénéité de la syntaxe des numéros ou nombres en tous genres | Dans le même fichier nous avons pour des chiffres parfois "1000,00" et parfois "100.000,00" | | | | | 5 | 1 | | | |
| 63 | 66 | Syntaxe | Chaîne alpha | Incohérences syntaxiques : l'usage du pluriel ou du singulier | | Rechercher les pluriels à l'aide de motifs d'expressions régulières (?) | | | | 5 | 1 | | | |
| 64 | 67 | Sémantique | Chaîne alpha | Plusieurs termes sont utilisés pour un même sens | Parfois on lit "Daesh", parfois "Isis" et parfois "Etat islamique" ; ou bien "agent" ou "commercial" ; etc. | Trier le champ concerné par ordre alphabétique et regarder les valeurs (?) | | | | 5 | | | | https://fr.wikipedia.org/wiki/R%C3%A9glementation_de_logiciel |
| 65 | 68 | Sémantique | Chaîne alpha | Certains termes sont mal régionalisés ou traduits dans la langue attendue | Dans un fichier où tout est en français, si l'on a "Grande-Bretagne" on devrait avoir "Etats-Unis" et pas "USA" qui est un terme anglais | Repérer les chaînes et lancer le correcteur d'orthographe dans la langue désirée (?) | | | | 5 | 1 | | | https://fr.wikipedia.org/wiki/R%C3%A9glementation_de_logiciel |
| 66 | 69 | Sémantique | Chaîne alpha | Certains termes, valeurs utilisées sont vieillis, inusités, cryptiques ou incompréhensibles | | Trier le champ concerné par ordre alphabétique et regarder les valeurs (?) | | | | 3 | 0 | | | |
| 67 | 70 | Sémantique | Chaîne alpha | Les abréviations ou sigles ne sont pas explicités | Wikipédia fournit des listes de très nombreux sigles : https://fr.wikipedia.org/wiki/Sigle | Trier le champ concerné par ordre alphabétique et regarder les valeurs (?) | | | | 3 | | | | |
| 68 | 71 | Sémantique | Chaîne alpha\|Nombre\|Date\|coordonnées | La valeur nulle est remplacée par une autre chaîne : zéro ou "-" ou "null" ou "1970-00-00" ou 0°00'00.0"N+0°00'00.0"E | 0°00'00.0"N+0°00'00.0"E est un problème car ce point existe mais il est placé en plein Atlantique | Trier le champ concerné par ordre alphabétique et regarder les valeurs (?) | | | | 3 | 1 | | | |
| 69 | 72 | Sémantique | Toutes | Inversion dans un couple de données | "Dupont Jean" au lieu de "Jean Dupont' | Repérer les couples de données et classer les colonnes par ordre alphabétique pour repérer une éventuelle inversion (?) | TODO : si une chaîne de la colonne_x est présent 3 fois dans la colonne_Y et inversement alors il y a suspicion d'inversion ? | | | 3 | 0 | | | Le producteur peut avoir saisi Prénom Nom en étant persuadé de cet ordre. Ce problème survient également pour des Prénom-Noms d'origine culturelle différente (les chinois utilisent Nom-Prénom) |
| 70 | 73 | Sémantique | Chaîne alpha | L'absence de lettres accentuées peut poser des problèmes de sens | "JUPE TUE LA FRANCE GAGNE" | Rechercher des colonnes alpha qui ne possèdent pas d'accent | egrep "[âêçéèëîôû]" file.csv \| wc -l | | | 1 | 1 | | | les accents sont signifiants en Français |
| 71 | 74 | Sémantique | Chaîne alpha\|Nombre | Erreur sémantique manifeste | Utilisation de "M" en lieu et place de "H" pour signifier un homme ; 69 pour le département en lieu et place du nom "Rhône" | Rechercher toutes les valeurs d'une colonne, les dédoublonner et analyser les résultats | csvcut -c colonne1 file.csv \| sort \| uniq | | | 5 | 0 | | | |
| 72 | 75 | Sémantique | Coordonnées | Erreur de système de coordonnées | Coordonnées en Lambert II au lieu de WGS 84 spécifié dans les métadonnées | | | | | 5 | | | | |
| 73 | 76 | Sémantique | Coordonnées | Les coordonnées géographiques sont données en degrés, minutes, secondes et non en degrés décimals, ce qui complique leur manipulation | 23°56'33" ou bien 23°56'33"E en lieu et place de la forme décimale 23,9756 | Regarder toutes les colonnes représentant des coordonnées | | | | 3 | 1 | | | |
| 74 | 77 | Sémantique | Date | Le format de la date est celui d'un autre pays ou d'une autre culture | 09/08/2016 au lieu de 08/09/2013 pour le 8 septembre 2016 (la syntaxe est correcte mais le sens est incorrect) | Rechercher toutes les valeurs d'une colonne, les dédoublonner, les trier et analyser les résultats | csvgrep -c colonne1 -r "\d?\d\/3\d" file.csv | | | 3 | 1 | | | |
| 75 | 78 | Sémantique | Chaîne alpha\|booléen | Liste de réponses fermée mal conçue : réponse "vrai" ou "faux" exclusivement alors que "sans réponse" ou autres pourraient convenir | "Vous êtes plutôt d'accord avec telle assertion : vrai-faux". "Ne se prononce pas" devrait pouvoir être une réponse pertinente. | | | | | 3 | 0 | | | |
| 76 | 79 | Sémantique | Chaîne alpha | Liste de réponses fermée mal conçue : présence de la réponse "Autre" ou "Divers" très fréquente | "Quel est votre ville favorite : Marseille, Paris, Autre" | Rechercher les chaînes "Autre" et "Divers" | perl -F/,/ -alne 'print if 1..1;print if m/Autre\|Divers\/i;' file.csv \| csvsort \| csvlook | | | 1 | 1 | | | Dans certains cas, les réponses "Autre" ou "Divers" peut être parfaitement justifiée. |
| 77 | 80 | Morpho-syntaxique | Toutes | Exprimer une donnée à travers un code difficile à manipuler | Mise en forme pour exprimer une donnée : couleur, gras, etc. | Parcourir visuellement l'ensemble du fichier | | | | 5 | 0 | | | |
| 78 | 81 | Morpho-syntaxique | Toutes | autres ? | Certains fichiers possèdent des cellules fusionnées : des données sont ajoutées sous forme de commentaires ; etc. | Parcourir visuellement l'ensemble du fichier | | | | 5 | | | | |
| 79 | 82 | Pertinence | Chaîne alpha\|Nombre | Aberration | * 197 ans (pour l'âge d'une personne)<br>* Général de Gaulle comme personne participant à un sondage | * Un classement des champs par ordre alphabétique permet de localiser des grandeurs aberrantes.<br>* Tester que les données vérifient la loi de Benford.<br>* WTFcsv. | | | | 30 | | | | |
| 80 | 83 | Pertinence | Chaîne alpha\|Nombre | Doute très raisonnable, valeurs inexplicables | 20 participants de plus de 110 ans | * Rechercher les valeurs extrêmes de chaque colonne et s'interroger.<br>* WTFcsv ? | | oui | | 5 | 1 | | | |
| 81 | 84 | Pertinence | Chaîne alpha | Certaines valeurs sont suspectes : 0000 ou xxxxxxxxxxxxx (à compléter) | - | | perl -F/,/ -alne 'print if 1..1;print if m/(000\|xxx)\;' file.csv \| csvsort \| csvlook | | | 3 | | | | |
| 82 | 85 | Pertinence | Nombre | Certaines valeurs sont suspectes : suites de chiffres comme 9999 ou 12345 | Des suites de 9999 ; nombreuses valeurs "12345" (détailler) | Recherche des chaînes "99*" et "123*" | perl -F/,/ -alne 'print if 1..1;print if m/999\|12345\|00\/\;' file.csv \| csvsort \| csvlook | | | 3 | 1 | | | |
| 83 | 86 | Pertinence | Date | Certaines valeurs sont suspectes : il existe des dates en 1900, 1904, 1969, 1970 | - | Recherche des chaînes "1900*", "1904*", "1969*", "1970*" | perl -F/,/ -alne 'print if 1..1;print if m/1900\|1904\|1969\|1970\;' file.csv \| csvsort \| csvlook | | | 3 | 1 | | | |
| 84 | 87 | Pertinence | Coordonnées | Certaines valeurs sont suspectes : il existe des coordonnées comme 0°00'00.0"N+0°00'00.0"E | 0°00'00.0"N+0°00'00.0"E est une valeur suspecte car c'est un point en plein milieu de l'Atlantique | POI : placer tous les POI sur une carte pour voir si certains sont hors périmètre | csvgrep -c colonne_x -r "(0°00'00).*" file.csv \| csvsort \| csvlook | | | 3 | 1 | | | |
| 85 | 88 | Pertinence | Toutes | La source n'est pas crédible (incompétent, juge et partie, etc.) | 15000 manifestants selon les organisateurs | Questionner la crédibilité de la source : est-elle compétente pour collecter ces données ? A-t-elle un intérêt partisan à faire parler les données dans une certaine direction ? | | oui | | 3 | 0 | | | Autres exemples : chiffres du chomage (?), chiffres "sortis du chapeau" par les politiques, résultats d'audiences ou financiers communiqués par l'acteur concerné par ces chiffres, ... |
| 86 | 89 | Pertinence | Toutes | Les données ont été hackées ou détournées | La source est crédible mais certains producteurs indirects ont pu agir pour que certaines données soient sur-représentées (sondage, etc.) | * La sur-représentation d'un profil ou des valeurs suspectes doivent conduire à s'interroger<br>* Tester que les données vérifient la loi de Benford. | | | | 30 | 0 | | | Exemple de l'affaire Clearstream. Affaire Hashley-Madison (?). Voir le type de problème "Le process d'acquisition n'est pas connu" |
| 87 | 90 | Réglementation | Toutes | Identification explicite de personnes sans déclaration CNIL | Prénom Nom ou numéro de tél. | | | oui | | | | | | |
| 88 | 91 | Réglementation | Toutes | Identification possible de personnes | Date et lieu de naissance | | | oui | | | | | | |
| 89 | 92 | Réglementation | Chaînes alpha | Il existe des jugements de valeurs à propos d'individus | "Client chiant", etc. | Rechercher des mots "interdits" comme "chiant", "stupide", "idiot", "connard/connasse", "encule", etc. | | | | 5 | 1 | | | |
| 90 | 93 | Réglementation | Chaînes alpha | Il existe des données de santé non anonymisées alors que les personnels qui les consultent n'y sont pas habilités | "Ne peut pas nous recevoir le mercredi matin car elle fait sa dialyse" | Rechercher des mots "interdits" comme "dialyse", "cancer", etc. | | | | 5 | 1 | | | |
| 91 | 94 | Réglementation | Toutes | Données d'origine ethnique ou relative à la religion des personnes | "Ne répond pas au téléphone le samedi (shabbat)" | Rechercher des mots qui peuvent être des indicateurs comme "caucasien", "chrétien", "juif", "musulman", etc. | | oui | | 3 | 0 | | | |
| 92 | 95 | Réglementation | Toutes | Données relatives aux opinions politiques, philosophiques ou à l'appartenance syndicale | "Lié au parti pirate" | Rechercher des mots qui peuvent être des indicateurs comme le nom de partis politiques, de courants de pensée, etc. | | oui | | 3 | 0 | | | |
| 93 | 96 | Réglementation | Toutes | Données relatives à la vie sexuelle ou au moeurs | "Ménage à 3" | Rechercher des mots qui peuvent être des indicateurs comme "sex", "homo", etc. | | oui | | 3 | 0 | | | |
| 94 | 97 | Réglementation | Toutes | Données tierces soumises à licence d'usage | Le fichier publié en Open Data utilise le géocodage de l'API de Google | Lire les métadonnées ; en cas de doute, demander explicitement au producteur. | | oui | | 5 | | | | |
| 95 | 98 | Réglementation | Chaîne alpha | Données relevant de la propriété littéraire et artistique sans autorisation d'usage : description textuelles | La description littéraire d'une chose est soumise à des droits | Rechercher les chaînes de plus 100 (?) caractère et évaluer si la rédaction dépasse un simple caractère factuel (?) | | | | 3 | | | | |
| 96 | 99 | Réglementation | binaire | Données relevant de la propriété littéraire et artistique sans autorisation d'usage : images ou fichiers multimédia | Les images d'une base de données sont soumises à des droits | Le jeu de données comprend-il des images ? Le droit d'usage de ces images est-il explicité ? Ce droit pose-t-il problème pour des usages ultérieurs ? | | | | 3 | | | | |
| 97 | 100 | Réglementation | Toutes | Données sensibles du point de vu de la sécurité des biens et des personnes | Plan d'une base militaire | Parcourir le fichier dans son ensemble suffit-il ? | | oui | | 3 | 0 | | | |
| 98 | 101 | Réglementation | Toutes | Données sensibles du point de vu de l'éthique | Localisation de minéraux rares ou de zones d'habitat d'espèces protégées | ? | | oui | | 3 | 0 | | | |
| 99 | 102 | Réglementation | Chaîne alpha\|Nombre | divers : exemple le capital social d'une entreprise s'exprime réglementairement arrondi à la valeur inférieure | | | | | | | 0 | | | Les chiffres communiqués aux impôts comme la TVA sont arrondis ; la raison sociale d'une entreprise ; les prénoms-noms dans un contexte d'identification officielle ; des tarifs (?) ; les cours des monnaies (?) ... |
| 100 | 103 | Manque | Nombre | La donnée est le résultat d'un calcul dont on n'a pas les données de départ | Le jeu de données contient un pourcentage, un rapport, une densité, etc. | Parcourir les métadonnées pour évaluer chaque champ | | | | 3 | 0 | | | |
| 101 | 104 | Manque | Toutes | Les trous : manque des "enregistrements" : des données dont vous connaissez l'existence sont manquantes | Il manque 10 communes dans la liste des mairies du département de la Savoie | Rechercher toutes les valeurs d'une colonne, les dédoublonner, les trier et analyser les résultats | | | | 3 | 0 | | | |
| 102 | 105 | Manque | Toutes | Les trous : manque des "enregistrements" : le tableau possède 65536 lignes | - | Ouvrir le jeu de données et regarder s'il contient 65536 lignes | csvstat --count file.csv<br><br>csvstat file.csv<br><br>wc -l file.csv # auquel il faut retirer 1 | | | 1 | 1 | | | |
| 103 | 106 | Manque | Toutes | Les trous : les données d'un champs sont tronquées | "10, av. du Général de Gau" est un exemple de champ tronqué à 25 caractères | * Pour une longueur de champ donné, quel pourcentage d'enregistrements remplissent ce champ complètement ? | csvgrep -c colonne_x -r ".{25}" file.csv \| csvlook<br><br>csvcut -c colonne_x file.csv \| csvgrep -c colonne_x -r ".{24}" \| csvsort -c adr \| uniq \| csvlook \| head -20 | | | 3 | 1 | | | |
| 104 | 107 | Manque | Toutes | Valeurs vides dans certains champs | - | Ouvrir le jeu de données et regarder s'il contient des valeurs vides | csvstat --null file.csv<br><br>csvstat file.csv | | | 1 | 1 | | | |
| 105 | 108 | Manque | Toutes | La granularité n'est pas suffisante | On a des pays, là où il serait intéressant d'avoir des régions ; on a des mètres là où certains usages nécessiteraient des cm | | | | | 5 | 0 | | | |
| 106 | 109 | Manque | Date | Le fuseau horaire n'est pas précisé dans un contexte de données réparties sur des fuseaux horaires différents | Pour une heure locale donnée, l'absence du fuseau horaire oblige le développeur a tenter de calculer l'heure GMT pour comparer des durées | | | | | 5 | 0 | | | |
| 107 | 110 | Manque | Toutes | L'insuffisance en matière de fréquence | L'état de feux tricolores classiques est donné tous les jours à minuit | Lire les métadonnées, examiner les données (un tri descendant des colonnes date peut aider) puis comparer avec des usages possibles | | | | 5 | 0 | | | |
| 108 | 111 | Manque | Toutes | L'insuffisance en matière de maillage | La pollution dans Paris est mesurée avec un seul capteur | | | | | 5 | 0 | | | |
| 109 | 112 | Manque | Toutes | L'insuffisance en matière de fraîcheur | Les chiffres du recensement de cette espèce date de 1976 | Lire les métadonnées, examiner les données (un tri descendant des colonnes date peut aider) puis comparer avec des usages possibles | | | | 5 | 0 | | | |
| 110 | 113 | Surabondance | Toutes | Doublons | * Jean Martin;21/12/1956;Luneville;noscontacts<br>* Jean Martin;21/12/1956;Luneville;listesuppl. | Repérer les champs permettant d'identifier une chose de manière non ambigue et analyser la présence de doublons à travers un tri alphabétique d'un des champs. | * Pour détecter les doublons stricts : sort file.xxx \| uniq -d<br>* Pour détecter les doublons selon deux colonnes : csvcut -c colonne1,colonne2 file.xxx \| sort \| uniq -d | | | 5 | 1 | | | La notion de doublons insclus les valeurs en triple, quadruple, etc. Il existe d'une part les doublons stricts, qui peuvent être fusionnés facilement ; d'autre par les doublons qui nécessitent une fusion des valeurs (plus difficiles à détecter et fusionner). |

| ID | N° d'ordre | Domaine du problème | Types de données concernées | Type de problème | Exemple | Détecter le problème | Ligne de commande pour le contrôle | Réduit considérablement l'usage des données | Niveau d'importance du problème | Temps d'analyse en min | Analyse semi-automatisable | Solution | Temps de traitement | Remarques |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 111 | 114 | Surabondance | Toutes | Valeur obtenue par calcul sur la base de deux données déjà présentes | Le milieu de deux coordonnées géographiques est indiqué en plus des deux coordonnées. | Lire les métadonnées et vérifier le calcul sur un petit échantillon | | | | 3 | 0 | | | |
| 112 | 115 | Surabondance | Toutes | Valeur renseignée alors qu'elle est calculable à partir de données prises ailleurs | Hors Paris et Marseille, pourquoi avoir un code postal quand la ville permet de le déduire ? | | | | | 3 | 0 | | | |
| 113 | 116 | Surabondance | Toutes | Surabondance de données en matière de : précision, fréquence, maillage ou fraicheur | La localisation au mm d'une porte d'entrée ; ajout du fuseau horaire dans contexte où toutes les données sont en France métropolitaine ; etc. | " La précision est-elle de plus de 4 chiffres après la virgule ? | csvgrep -c colonne_x -r ".*\d{\,}\.\d\d\d\d+" file.csv \| csvsort \| csvlook | | | 3 | 1 | | | |
| 114 | 117 | Surabondance | Coordonnées | Pour des coordonnées géographiques, une précision supérieure à 8 unités après la virgule est inutile (précision de l'ordre du mm) ; 5 chiffres après la virgule donnent déjà une précision de l'ordre du mètre | 23,73825619 positionne un objet à environ 1 mm | Regarder la longueur des champs représentant des coordonnées | csvgrep -c colonne_x -r ".*\d{\,}\.\d\d\d\d\d\d\d+" file.csv \| csvsort \| csvlook | | | 3 | 1 | Arrondir à 7 chiffre après la virgule, voire à 6 chiffres. | | |
| | | | | | | | | | | | | | | |
| | | | | | | | 25 | | | 425 | 39 | 26 | | |

| Sprint qualité | Sprint qualité basé sur la version 0.2 de la méthodo "sprint qualité" : http://infolabs.io/sprint-qualite |
|---|---|
| **Titre initial du jeu de données** | Base de données accidents corporels de la circulation |
| URL de référence | https://www.data.gouv.fr/fr/datasets/base-de-donnees-accidents-corporels-de-la-circulation/ |
| Date du sprint (aaaa-mm-jj) | 10/11/2016 |
| Auteurs<br><br>Binôme 1<br>Prénom-Nom <email>,<br>Prénom2-Nom2 <email2><br><br>Binôme 2<br>Prénom-Nom <email>,<br>Prénom2-Nom2 <email2><br><br>etc. | Animateurs : Armelle Gilliard <agilliard@fing.org>, Charles Nepote <charles.nepote@fing.org><br><br>Binôme 1<br>Romain Talès<br>Christian Quest<br><br>Binôme 2<br>Marie Heuze<br>Vincent Bataille<br><br>Binôme 3<br>Sarah Labelle<br>Fabien Antoine |
| Remarques | On étudie le fichier caracteristiques_2015.csv |
| URL de publication du résultat du sprint | |

| ID | N° | Type de problème | Exemple | Observateurs | Problème rencontré ? | Commentaire | Exemple réel dans le jeu de donnée | A quel niveau ou fréquence le problème est-il observé ? [1] | Impact du problème sur les usages [2] | Effort de correction estimé | Méthode ou process de correction rejouable | Correcteurs | Complétude de la correction [3] | Date de la correction | Emplacement du nouveau jeu de données corrigé (URL, localisation, etc.) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | Le jeu de données est dans un format "image" ne permettant pas de manipuler les données | Le jeu de données est un fichier image au format JPEG ou PDF. | Animateur | non | | - | | | | | | | | |
| 2 | 2 | Le jeu de données est dans un format non spécifiquement adapté aux données : PDF, Word, ODF, epub, HTML, SVG, etc. | Le jeu de données est un fichier HTML. | Animateur | non | | - | | | | | | | | |
| 3 | 3 | Le format du jeu de données n'est pas précisé (fichier CSV, TSV, etc.) | l'extension du jeu de données ne permet pas de savoir quel logiciel permet de l'ouvrir et l'éditeur n'a pas fourni d'indication complémentaire | Animateur | non | | - | | | | | | | | |
| 31 | 4 | La documentation et les métadonnées sont quasi inexistantes voire absentes | La documentation tient sur 5 lignes alors que le fichier est très complexe | Binôme 2 | non | OK | | | | | | | | | |
| 115 | 5 | La documentation et les métadonnées sont d'un usage difficile (doc papier, doc au format PDF image, doc uniquement en anglais, etc.) | La documentation est fourni sous forme de PDF image : les usagers ne peuvent pas rechercher des termes pour y naviguer rapidement | Animateur | non | OK | | | | | | | | | |
| 4 | 6 | La licence du jeu de données ne nous permet pas de l'utiliser | Le jeu de données est un fichier commercial que l'on n'a pas acheté | Binôme 1 | non | définir le "nous" | | | | | | | | | |
| 5 | 7 | Le format du jeu de données n'est pas ouvert | Le fichier est au format .xls ou .xlsx | Binôme 1 | non | | | | | | | | | | |
| 6 | 8 | Le format du jeu de données ne permet pas d'ouvrir le fichier dans des outils très répandus (Excel, Notepad…) | Le fichier au format .csv s'ouvre mal dans Excel, outil le plus répandu pour ouvrir des tableaux | Binôme 1 | oui | fichier csv pas simple d'ouverture avec Excel | | 1. concerne l'ensemble du jeu | | | | | | | |
| 7 | 9 | L'encodage du fichier n'est pas spécifié (ISO-8859-1, UTF8, etc.) | Le fichier contient des caractères ésotériques mais on ne sait pas s'il s'agit d'un problème d'encodage | Binôme 3 | oui | fichier au format ISO-8859-1 | non spécifié | 1. concerne l'ensemble du jeu | faible | faible | | | | | |
| 8 | 10 | L'encodage n'est pas en UTF-8 ; ce dernier devient la norme de facto et d'autres encodages peuvent engendrer des problèmes | L'encodage est en ISO-8859-1 | Binôme 1 | oui | problème d'encodage non UTF8 | pbl avec les caractères accentués | 1. concerne l'ensemble du jeu | faible | très faible | iconv -f ISO-8859-1 -t UTF-8 ./caracteristiques_2015.csv > caracteristiques_2015_rev1.csv | | | | |
| 9 | 11 | L'encodage n'est pas homogène | Certaines données sont correctement encodées et d'autres contiennent des caractères ésotériques | Animateur | oui | | "ç?n?al de Gaulle", "poincaŽ?" au côté de "général" ou "poincaré" ; voir caractère bizare ligne "201500013582" | 4. problème fréquent | faible | important | | | | | |
| 10 | 12 | Le fichier est mal formé | Pour certaines lignes, parfois une colonne manque, ou, le fichier cvs comporte des "virgule" non formatées et empêche l'ouverture correcte du fichier | Binôme 1 | non | | | | | | | | | | |
| 11 | 13 | Le jeu de données concernant des horaires de mode de transport ne possède pas de version au format GTFS | Le fichier n'est pas au format GTFS | Animateur | sans objet | | | | | | | | | | |
| 12 | 14 | Le jeu de données concernant des œuvres n'est pas au format Dublin Core | Le fichier n'est pas au format Dublin Core | Animateur | sans objet | | | | | | | | | | |
| 13 | 15 | Le jeu de données utilise une norme peu accessible au plus grand nombre (coût, complexité) | Le jeu de données est au format TRIDENT | Animateur | sans objet | | | | | | | | | | |
| 14 | 16 | Le process d'acquisition n'est pas connu | Wiki leaks | Binôme 1 | non | | | | | | | | | | |
| 15 | 17 | L'échantillon n'est pas documenté | L'échantillon semble représentatif mais on ne peut pas vérifier qu'il le soit bien, puisque ce dernier n'est pas documenté | Binôme 1 | non | aucun problème | | | | | | | | | |
| 16 | 18 | Le format d'un des champs n'est pas documenté, si bien qu'on ne peut comprendre ce qu'il contient ou bien contrôler ses valeurs | * La date est parfois exprimée par le nombre de secondes depuis 1970 ; cette donnée est difficile à comprendre.<br>* Un jeu de données contient un champ "image" en binaire, dont le format n'est pas spécifié. | Binôme 1 | non | bien documenté | | | | | | | | | |
| 17 | 19 | La taille maximale d'un champ n'est pas documentée | On ne sait pas si un code peu dépasser 10 caractères et si certaines valeurs sont donc fausses | | | | | | | | | | | | |
| 18 | 20 | Pour tel champ, l'incertitude de la mesure n'est pas documentée (appelée aussi "précision", exprimée en % ou bien "à plus plus ou moins X unités près") | Des coordonnées GPS sont indiquées mais on ne connaît pas leur marge d'erreur (précises à 10 m, à 100 m ?) ; la précision d'une mesure de température n'est pas explicitée (+/- 0,1° ? +/- 1° ?) | Binôme 2 | oui | KO: Non précisée sur les lat/lon | | 1. concerne l'ensemble du jeu | important | important | | | | | |
| 117 | 21 | L'incertitude de la mesure n'est pas connue par le producteur | Le producteur des données ne connaît pas la précision de ses mesures | | | | | | | | | | | | |
| 22 | 22 | La précision n'est pas cohérente avec la granularité : l'incertitude de la mesure est 100 fois supérieure à la granularité | Des coordonnées géographiques annoncent une granularité au cm alors que l'incertitude des appareils de mesure est de +/- 5 mètres | | | | | | | | | | | | |
| 19 | 23 | L'origine de certaines données est une entrée manuelle non contrôlée | Le risque est d'obtenir 25 orthographes de "Saint-André-des-Arts" | Binôme 1 | oui | aucune information sur les saisies contrôlées | La lecture du champ "adr" suggère une entrée manuelle tant la disparité des formats d'adresse est importante | 1. concerne l'ensemble du jeu | important | important | | | | | |
| 20 | 24 | Les données proviennent d'un processus de reconnaissance automatique dont la marge d'erreur est globalement bonne mais localement problématique (OCR, reconnaissance de forme, géocodage, etc.) | OCR ;<br>reconnaissance automatique des visages (va dépendre de la qualité de la lumière de la prise de vue, de la couleur des personnes concernées (c'est encore un problème en 2016)) ; etc. | Binôme 1 | oui | la question se pose pour les coordonnées GPS. Il n'est pas indiqué si elles sont mesurées sur le terrain ou par géocodage | | 1. concerne l'ensemble du jeu | important | important | | | | | |
| 21 | 25 | L'échantillon est biaisé | Certaines populations sont absentes, sur-représentées ou sous-représentées ; les données subissent une forte variation saisonière | Binôme 1 | non | | | | | | | | | | |
| 23 | 26 | Le process de signalement d'erreur et d'échange avec le producteur n'est pas explicité | Aucune forme de contact n'est donnée | Binôme 3 | oui | le site data.gouv fournit un processus de signalement, mais le producteur n'est pas directement impliqué | | 1. concerne l'ensemble du jeu | faible | faible | | | | | |
| 24 | 27 | Le process de signalement d'erreur et d'échange avec le producteur n'existe pas ou bien il est défaillant | Le producteur ne répond pas aux questions | Binôme 2 | oui | OK Un email générique dans lefooter de la doc<br>KO Pas de réponses sur data.gouv.fr | | 1. concerne l'ensemble du jeu | faible | faible | | | | | |
| 25 | 28 | La disponibilité de la donnée n'est pas documentée (temps pendant lequel la donnée est accessible par rapport au temps total souhaité, généralement exprimé en pourcentage) | L'utilisateur ne sait pas si la qualité de service est de 95% ou 99,99%. Si le système qui héberge la donnée est régulièrement inaccessible (maintenance, etc.), les usagers devraient en être informé pour savoir si leur usage en est impacté | Binôme 3 | oui | | | 1. concerne l'ensemble du jeu | nul ou négligeable | important | | | | | |
| 26 | 29 | La disponibilité de la donnée n'est pas mesurée | Le producteur ne sait pas si la qualité de service est de 95% ou 99.99 alors que tel futur usage est critique | Binôme 2 | | On ne sait pas | | | | | | | | | |
| 116 | 30 | Le mode d'accès à la donnée est un frein à l'usage (temps d'accès, droit d'accès long et complexe, droit d'accès limité) | La requête d'une donnée "temps réel" met plus de 40 secondes ; l'accès à la donnée nécessite un certificat de sécurité long à obtenir ; l'architecture du site ne permet pas à un robot de télécharger les actualisations des données | | | | | | | | | | | | |
| 27 | 31 | La mesure de la qualité n'est pas documentée | Des contrôles qualité existent (amont ou aval) mais ils ne sont pas explicités si bien qu'on ne peut savoir si tel champ est fiable ou non | Binôme 3 | oui | ce n'est pas documenté | la qualité des différents champs n'est pas décrite. la description n'est pas adjointe à une étude sur la fiabilité de ces champs, notamment sur les lieux, communes, adresses, gps. | 1. concerne l'ensemble du jeu | important | important | | | | | |
| 28 | 32 | La qualité de la donnée n'est pas mesurable à travers des contrôles formels | Il n'existe pas de méthode de contrôle permettant de dire si la syntaxe de ce champ est bonne | Binôme 2 | oui | KO ??? | | 1. concerne l'ensemble du jeu | faible | important | | | | | |
| 29 | 33 | La qualité de la données n'est pas mesurée | Aucune méthode de contrôle n'est mise en oeuvre pour mesurer la qualité des données | | | | | | | | | | | | |
| 30 | 34 | Une entité possède plusieurs identifiants | | Binôme 2 | non | OK | | | | | | | | | |
| 32 | 35 | Le nom ou titre du jeu de données est vague, ambigu ou trop complexe : titre de la notice éditoriale, nom donné dans les métadonnées ou dans la documentation (pas le nom du fichier) | * "Résultat des élections" : lesquelles ? où ? quand ?<br>* "Résultats des élections à Montréal" : il existe 6 communes appelées Montréal dans le monde.. | Binôme 2 | non | On pourrait préciser "en France (2006-2015)" | Jeu de données sur data.gou.fr : OK<br>Nom de la ressource : KO<br>Nom du fichier de documentation : KO | | | | | | | | |
| 33 | 36 | Manque de métadonnées : fourchette temporelle non explicitée | Des dates figurent dans le jeu mais aucune métadonnée ne peut confirmer la fourchette attendue de ces dates. Exemple : Trésorerie du 01/02/2010 au 24/11/2016. | Binôme 2 | non | OK | | | | | | | | | |
| 34 | 37 | Manque de métadonnées : zone spatiale non explicitée | Des coordonnées figurent dans le jeu mais aucune métadonnée ne peut confirmer la zone d' appartenance attendue pour ces points. | Binôme 2 | non | OK | | | | | | | | | |
| 35 | 38 | Manque de métadonnées : fourchette non spécifiée | On peut attendre d'un nombre qu'il soit compris entre une valeur minimum et une valeur maximum ; par exemple l'âge d'une personne devrait toujours être entre 0 et 130 voir 18 et 70 selon les cas. | Binôme 2 | non | OK | | | | | | | | | |
| 36 | 39 | Manque de métadonnées : le fait que le champ soit un booléen n'est pas spécifié | | Binôme 2 | sans objet | | | | | | | | | | |
| 37 | 40 | Manque de métadonnées : le format du booléen n'est pas spécifié | On ne sait pas à quelles valeurs s'attendre : "vrai"-" faux" ou "oui"-"non" | Binôme 2 | sans objet | | | | | | | | | | |
| 38 | 41 | Manque de métadonnées : processus et contexte de production non explicités | On ne sait pas si une mesure vient d'un capteur ou d'une mesure manuelle | Binôme 3 | oui | document ne décrit pas le processus | (forces de l'ordre sur le terrain qui enregistre les données- priorité 3 dans leur travail) | 1. concerne l'ensemble du jeu | compréhension des conditions de renseignement sur les caractéristiques de l'accident | faible | | | | | |
| 39 | 42 | Manque de métadonnées : la fraicheur des données n'est pas explicité :<br>* le délai entre le réel et la mise en base de la donnée<br>* le délai entre le réel et la publication de la donnée | Il n'est pas dit si telle information sur une grossesse va mettre plus de neuf avant d'arriver au réutilisateur | Binôme 3 | oui | | fraîcheur décrite sur deux plans : par objet et pour la base | 1. concerne l'ensemble du jeu | faible | important | | | | | |
| 40 | 43 | Manque de métadonnées : la langue des textes n'est pas spécifiée | | Binôme 3 | oui | spécifié dans la boîte data.gouv | | 1. concerne l'ensemble du jeu | nul ou négligeable | faible | | | | | |
| 41 | 44 | Métadonnées imprécises : le format de date n'est pas spécifié | Format américain ? anglais ? européen ? etc. | Binôme 3 | oui | le format n'est pas spécifié | dans la doc le format numérique de la date n'est pas précisé | 1. concerne l'ensemble du jeu | faible | faible | | | | | |
| 42 | 45 | Métadonnées imprécises : unités non spécifiées | On ne dit pas si colonne "hauteur" est en cm ou dm | Binôme 3 | non | pas de données numériques hors données dates et gps | | | | | | | | | |
| 43 | 46 | Métadonnées imprécises : système de coordonnées non spécifié | La documentation n'indique pas si les coordonnées sont en WGS 84, Lambert ou un autre système | Binôme 3 | oui | l'unité GPS n'est qu'imparfaitement décrite, | dans la documentation il n'y a pas d'exemple sur les coordonnées GPS | 1. concerne l'ensemble du jeu | important | faible | | | | | |
| 44 | 47 | Métadonnées imprécises : nom de colonnes ambigus | "Emplacement" ne dit rien sur la donnée attendue : une adresse ? "en haut" ? "devant" ? etc. | Binôme 3 | oui | oui c'est ambigü, il est nécessaire d'ouvrir la documentation pour comprendre les champs | ex: abréviation "col" pour "type de collision" ou "int" pour "intersection" | 1. concerne l'ensemble du jeu | faible | important | | | | | |
| 45 | 48 | Métadonnées fausses | | Binôme 3 | | difficile à voir sans ouvrir le jeu de données | | | | | | | | | |
| 46 | 49 | La taille maximale d'un champ dépasse celle qui est spécifiée dans la documentation | La colonne "âge" spécifie une longueur de 3 caractères maximum et certaines valeurs sont des 4 caractères ou plus | | | | | | | | | | | | |
| 47 | 50 | L'ordre des colonnes ne correspond pas à l'ordre donné dans la documentation | La documentation donne Prénom;Nom;Âge;Profession alors que le jeu se présente sous la forme Nom;Prénom;Âge;Profession | | | | | | | | | | | | |
| 48 | 51 | Les coordonnées ne sont pas au format WGS 84 | Les coordonnées sont au format Lambert II nécessitant une conversion des points pour des usages mobile liés à des GPS grand public | | | | | | | | | | | | |
| 49 | 52 | Les codes pays ne sont pas au format ISO 3166 | L'Allemagne est noté "ALL" alors qu'il existe un code ISO employé internationalement | Animateur | sans objet | | | | | | | | | | |
| 50 | 53 | Les codes de langues ne sont pas au format ISO 639 | Le français est noté "F" ou "français" en lieu et place de "fr" | Animateur | sans objet | | | | | | | | | | |
| 51 | 54 | Les dates ne sont pas au format ISO 8601 | La date est notée "01/01/2016" | Animateur | oui | | | 1. concerne l'ensemble du jeu | faible | faible | | | | | |
| 52 | 55 | Les monnaies ne sont pas au format ISO 4217 | Le franc suisse est noté FS | Animateur | sans objet | | | | | | | | | | |
| 53 | 56 | Erreurs syntaxiques : espace(s) au début ou à la fin du champ | " Pierre" au lieu de "Pierre" | | | | | | | | | | | | |
| 54 | 57 | Erreurs syntaxiques : bug syntaxique dans les strates du SI : le cas de l'apostrophe | "N'Diaye" à la place de N'Diaye | | | | | | | | | | | | |
| 55 | 58 | Erreurs syntaxiques : syntaxe des numéros ou nombres en tous genres | "456" au lieu de "45"; "1,000,000" au lieu de "1000000" | | | | | | | | | | | | |
| 56 | 59 | Erreurs syntaxiques : codes (code INSEE, code postal, SIRET, SIREN, n° de Sécu, ISBN, ISSN, IBAN, BIC, code ROM, indicatif du pays, code APE, code NAF, etc.) | 7100 au lieu de 07100 pour un code postal | Binôme 1 | non | dep OK, com OK | | | | | | | | | |
| 57 | 60 | Erreurs syntaxiques : sigles et abréviations | "SNCF", "S. N. C. F.", "S.C.N.F." ? "Boul" ou "Boul." ou "Bld" ? | | | | | | | | | | | | |
| 58 | 61 | Erreurs syntaxiques : booléen | "V" au lieu de "1" selon la spécification du booléen | Binôme 1 | sans objet | pas de booléen dans le jeu de données | | | | | | | | | |
| 59 | 62 | Erreurs syntaxiques : email, url | laurent.dupont@wanadoo@fr | Binôme 1 | sans objet | | | | | | | | | | |
| 60 | 63 | Erreur syntaxique sur la date | 2016/09/30 au lieu de 2016-09-30 attendu | Binôme 3 | oui | le format n'est pas spécifié, les 00 significatifs des dates ne sont pas rentrés | la date et l'heure sont indiquées pour 00:45 le 22 novembre 2015 est écrit 15.11.22.45 | 1. concerne l'ensemble du jeu | faible | faible | | | | | |
| 61 | 64 | Incohérences syntaxiques : syntaxe des noms propres | "de La Tour" ou "La Tour (de)" ? | Binôme 3 | oui | seul le champ adresses est concerné | deux modalités de rédaction : rue GHESQUIERE et rue Ghesquière sur la commune 59052 (Bauvin) | 5. problème très fréquent | faible | important | | | | | |
| 62 | 65 | Incohérences syntaxiques : homogénéité de la syntaxe des numéros ou nombres en tous genres | Dans le même fichier nous avons pour des chiffres parfois "1000.00" et parfois "100.000.00" | Binôme 3 | oui | cohérence hors gps et date | | | | | | | | | |
| 63 | 66 | Incohérences syntaxiques : l'usage du pluriel ou du singulier | | Binôme 3 | sans objet | | | | | | | | | | |
| 64 | 67 | Plusieurs termes sont utilisés pour un même sens | Parfois on lit "Daesh", parfois "Isis" et parfois "EI" ; ou bien "agent" ou "commercial", etc. | Animateur | sans objet | | | | | | | | | | |
| 65 | 68 | Certains termes sont mal régionalisés ou traduits dans la langue attendue | Dans un fichier où tout est en français, si l'on a "Grande-Bretagne" on devrait avoir "Etats-Unis" et pas "USA" qui est un terme anglais | Binôme 3 | non | | | | | | | | | | |
| 66 | 69 | Certains termes, valeurs utilisées sont vieillis, inusités, cryptiques ou incompréhensibles | | | | | | | | | | | | | |
| 67 | 70 | Les abréviations ou sigles ne sont pas explicités | Wikipédia fournit des listes de très nombreux sigles : https://fr.wikipedia.org/wiki/Sigle | | | | | | | | | | | | |

| ID | N° | Type de problème | Exemple | Observateurs | Problème rencontré ? | Commentaire | Exemple réel dans le jeu de donnée | A quel niveau ou fréquence le problème est-il observé ? [1] | Impact du problème sur les usages [2] | Effort de correction estimé | Méthode ou process de correction rejouable | Correcteurs | Complétude de la correction [3] | Date de la correction | Emplacement du nouveau jeu de données corrigé (URL, localisation, etc.) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 68 | 71 | La valeur nulle est remplacée par une autre chaîne : zéro ou "-" ou "null" ou "1970-00-00" ou 0°00'00.0"N+0° 00'00.0"E | 0°00'00.0"N+0°00'00.0"E est un problème car ce point existe mais il est placé en plein Atlantique | Binôme 2 | oui | Les latlon manquantes sont à 00000 au lieu de null | | 4. problème fréquent | | | | | | | |
| 69 | 72 | Inversion dans un couple de données | "Dupont Jean" au lieu de "Jean Dupont" | Binôme 2 | non | OK : Pas d'inversion évidente latlon ou hh/mm | | | | | | | | | |
| 70 | 73 | L'absence de lettres accentuées peut poser des problèmes de sens | "JUPE TUE LA FRANCE GAGNE" | Binôme 2 | oui | cap non acccentuées sur nombre des adresses | | 4. problème fréquent | | | | | | | |
| 71 | 74 | Erreur sémantique manifeste | Utilisation de "M" en lieu et place de "H" pour signifier un homme ; 69 pour le département en lieu et place du nom "Rhône" | | | | | | | | | | | | |
| 72 | 75 | Erreur de système de coordonnées | Coordonnées en Lambert II au lieu de WGS 84 spécifié dans les métadonnées | Binôme 2 | non | | | | | | | | | | |
| 73 | 76 | Les coordonnées géographiques sont données en degrés, minutes, secondes et non en degrés décimals, ce qui complique leur manipulation | 23°56'33" ou bien 23°56'33"E en lieu et place de la forme décimale 23,9756 | Animateur | oui | Pas de décimale dans les degrés décimaux des latlon | | 5. problème très fréquent | | | | | | | |
| 74 | 77 | Le format de la date est celui d'un autre pays ou d'une autre culture | 09/08/2016 au lieu de 08/09/2013 pour le 8 septembre 2016 (la syntaxe est correcte mais le sens est incorrect) | Binôme 2 | non | | | | | | | | | | |
| 75 | 78 | Liste de réponses fermée mal conçue : réponse "vrai" ou "faux" exclusivement alors que "sans réponse" ou autres pourraient convenir | "Vous êtes plutôt d'accord avec telle assertion : vrai-faux". "Ne se prononce pas" devrait pouvoir être une réponse pertinente. | Binôme 2 | sans objet | OK | | | | | | | | | |
| 76 | 79 | Liste de réponses fermée mal conçue : présence de la réponse "Autre" ou "Divers" très fréquente | "Quel est votre ville favorite : Marseille, Paris, Autre" | Binôme 2 | oui | Ily a des "Autres" mais est-ce vraiment un problème ? | | | | | | | | | |
| 77 | 80 | Exprimer une donnée à travers un code difficile à manipuler | Mise en forme pour exprimer une donnée : couleur, gras, etc. | Binôme 1 | oui | Utilisation de code non conforme pour la Corse (2012/02) au lieu du code standard 2A/2B | | 4. problème fréquent | important | faible | | | | | |
| 78 | 81 | autres ? | Certains fichiers possèdent des cellules fusionnées ; des données sont ajoutées sous forme de commentaires ; etc. | | | | | | | | | | | | |
| 79 | 82 | Aberration | * 197 ans (pour l'âge d'une personne)<br>* Général de Gaulle comme personne participant à un sondage | | | | | | | | | | | | |
| 80 | 83 | Doute très raisonnable, valeurs inexplicables | 20 participants de plus de 110 ans | | | | | | | | | | | | |
| 81 | 84 | Certaines valeurs sont suspectes : 0000 ou xxxxxxxxxxxx (à compléter) | - | | | | | | | | | | | | |
| 82 | 85 | Certaines valeurs sont suspectes : suites de chiffres comme 9999 ou 12345 | Des suites de 9999 ; nombreuses valeurs "12345" (détailler) | | | | | | | | | | | | |
| 83 | 86 | Certaines valeurs sont suspectes : il existe des dates en 1900, 1904, 1969, 1970 | - | | | | | | | | | | | | |
| 84 | 87 | Certaines valeurs sont suspectes : il existe des coordonnées comme 0° 00'00.0"N+0°00'00.0"E | 0°00'00.0"N+0°00'00.0"E est une valeur suspecte car c'est un point en plein milieu de l'Atlantique | | | | | | | | | | | | |
| 85 | 88 | La source n'est pas crédible (incompétent, juge et partie, etc.) | 15000 manifestants selon les organisateurs | | | | | | | | | | | | |
| 86 | 89 | Les données ont été hackées ou détournées | La source est crédible mais certains producteurs indirects ont pu agir pour que certaines données soient sur-représentées (sondage, etc.) | | | | | | | | | | | | |
| 87 | 90 | Identification explicite de personnes sans déclaration CNIL | Prénom Nom ou numéro de tél. | | | | | | | | | | | | |
| 88 | 91 | Identification possible de personnes | Date et lieu de naissance | | | | | | | | | | | | |
| 89 | 92 | Il existe des jugements de valeurs à propos d'individus | "Client chiant", etc. | | | | | | | | | | | | |
| 90 | 93 | Il existe des données de santé non anonymisées alors que les personnels qui les consultent n'y sont pas habilités | "Ne peut pas nous recevoir le mercredi matin car elle fait sa dialyse" | | | | | | | | | | | | |
| 91 | 94 | Données d'origine ethnique ou relative à la religion des personnes | "Ne répond pas au téléphone le samedi (shabbat)" | | | | | | | | | | | | |
| 92 | 95 | Données relatives aux opinions politiques, philosophiques ou à l'appartenance syndicale | "Lié au parti pirate" | | | | | | | | | | | | |
| 93 | 96 | Données relatives à la vie sexuelle ou au moeurs | "Ménage à 3" | | | | | | | | | | | | |
| 94 | 97 | Données tierces soumises à licence d'usage | Le fichier publié en Open Data utilise le géocodage de l'API de Google | | | | | | | | | | | | |
| 95 | 98 | Données relevant de la propriété littéraire et artistique sans autorisation d'usage : description textuelles | La description littéraire d'une chose est soumise à des droits | | | | | | | | | | | | |
| 96 | 99 | Données relevant de la propriété littéraire et artistique sans autorisation d'usage : images ou fichiers multimédia | Les images d'une base de données sont soumises à des droits | | | | | | | | | | | | |
| 97 | 100 | Données sensibles du point de vu de la sécurité des biens et des personnes | Plan d'une base militaire | | | | | | | | | | | | |
| 98 | 101 | Données sensibles du point de vu de l'éthique | Localisation de minéraux rares ou de zones d'habitat d'espèces protégées | | | | | | | | | | | | |
| 99 | 102 | divers : exemple le capital social d'une entreprise s'exprime réglementairement arrondi à la valeur inférieure | | | | | | | | | | | | | |
| 100 | 103 | La donnée est le résultat d'un calcul dont on n'a pas les données de départ | Le jeu de données contient un pourcentage, un rapport, une densité, etc. | | | | | | | | | | | | |
| 101 | 104 | Les trous : manque des "enregistrements" : des données dont vous connaissez l'existence sont manquantes | Il manque 10 communes dans la liste des mairies du département de la Savoie | | | | | | | | | | | | |
| 102 | 105 | Les trous : manque des "enregistrements" : le tableau possède 65536 lignes | - | | | | | | | | | | | | |
| 103 | 106 | Les trous : les données d'un champs sont tronquées | "10, av. du Général de Gau" est un exemple de champ tronqué à 25 caractères | | | | | | | | | | | | |
| 104 | 107 | Valeurs vides dans certains champs | - | | | | | | | | | | | | |
| 105 | 108 | La granularité n'est pas suffisante | On a des pays, là où il serait intéressant d'avoir des régions ; on a des mètres là où certains usages nécessiteraient des cm | | | | | | | | | | | | |
| 106 | 109 | Le fuseau horaire n'est pas précisé dans un contexte de données réparties sur des fuseaux horaires différents | Pour une heure locale donnée, l'absence du fuseau horaire oblige le développeur a tenter de calculer l'heure GMT pour comparer des durées | Binôme 2 | oui | KO pas de précision de timezone alors que certains accidents ont lieu dans les DOM | | 1. concerne l'ensemble du jeu | faible | faible | | | | | |
| 107 | 110 | L'insuffisance en matière de fréquence | L'état de feux tricolores classiques est donné tous les jours à minuit | | | | | | | | | | | | |
| 108 | 111 | L'insuffisance en matière de maillage | La pollution dans Paris est mesurée avec un seul capteur | | | | | | | | | | | | |
| 109 | 112 | L'insuffisance en matière de fraîcheur | Les chiffres du recensement de cette espèce date de 1976 | | | | | | | | | | | | |
| 110 | 113 | Doublons | * Jean Martin;21/12/1956;Lunéville;noscontacts<br>* Jean Martin;21/12/1956;Lunéville;listesuppl. | | | | | | | | | | | | |
| 111 | 114 | Valeur obtenue par calcul sur la base de deux données déjà présentes | Le milieu de deux coordonnées géographiques est indiqué en plus des deux coordonnées. | | | | | | | | | | | | |
| 112 | 115 | Valeur renseignée alors qu'elle est calculable à partir de données prises ailleurs | Hors Paris et Marseille, pourquoi avoir un code postal quand la ville permet de le déduire ? | | | | | | | | | | | | |
| 113 | 116 | Surabondance de données en matière de : précision, fréquence, maillage ou fraîcheur | La localisation au mm d'une porte d'entrée ; ajout du fuseau horaire dans contexte où toutes les données sont en France métropolitaine ; etc. | | | | | | | | | | | | |
| 114 | 117 | Pour des coordonnées géographiques, une précision supérieure à 8 unités après la virgule est inutile (précision de l'ordre du mm) ; 5 chiffres après la virgule donnent déjà une précision de l'ordre du mètre | 23,73825619 positionne un objet à environ 1 mm | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | |
| | | Nombre de problèmes détectés | 28 | | | | | | | | | | | | |
| | | Nombre de problèmes contrôlés | 64 | | | | | | | | | | | | |
| | | moyenne des problèmes analysés par binôme | 21,33333333 | | | | | | | | | | | | |
| | | Début de la recherche : 10h10<br>Fin de la recherche : 11h40 (1h30=90 min) | 90 | | | | | | | | | | | | |
| | | Durée moyenne d'analyse d'un problème | 4,21875 | | | | | | | | | | | | |

[1] * très faible
* faible
* importante
* très importante

[2] * négligeable
* faible
* important
* rend le jeu inutilisable

[3] Complète
Partielle