II. — Qualité des données, les grands types de problèmes

Cette partie va s'intéresser aux **problèmes pratiques** de la qualité des données, sous trois angles :

- **quels sont-ils ?** Nous les décrirons, notamment sous l'angle de leurs impacts concrets.
- **comment les détecter ?** Nous regarderons notamment les cas où l'outillage voire l'automatisation de la détection est possible.
- **comment les corriger ?** Nous donnerons ici des pistes d'actions, sachant que cet aspect est très dépendant du contexte d'utilisation.

Pour ce faire nous avons élaboré une liste de **120 points de contrôle** et les avons classé par grandes catégories. Ces 120 points concourent à la méthodologie du **Sprint qualitét**, conçue et réalisée par la Fing, qui offre un pendant opérationnel à ce guide : http://infolabs.io/sprint-qualite

Chaque point de contrôle fait l'objet d'un identifiant que nous réutilisons dans ces différents documents, afin d'en faciliter l'usage.

Tous les points de contrôle ne sont pas détaillés, soit qu'il n'en offre pas l'intérêt, soit que nous n'en ayons pas encore eu le temps. Ce document fera l'objet d'une évolution en 2018.

1. L'environnement des données : fichier(s), encodage, métadonnées

L'environnement des données est un des domaines les plus importants de la qualité des données.

Les données elles-mêmes peuvent être d'une qualité irréprochable, mais si le fichier, l'encodage ou les métadonnées sont mal ou peu (voire pas !) documentées, les données peuvent devenir **inutilisables**. Nous avons identifié **plus d'une quarantaine de problématiques** liées à l'environnement des données. Certaines ont un impact si fort qu'elles peuvent rendre l'usage impossible.

Prenons un exemple simple :
"ID25 : *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)*".
On comprend tout de suite que sans un minimum d'information sur cette question, l'usage de cette donnée sera impossible dans des cas courants : la nécessité d'une information en temps réel, un système d'alerte critique, etc.

Les problèmes d'accessibilité aux données

Ces points de contrôle arrivent très tôt dans ce document car ils posent des problèmes concernant l'accès même aux données. Nous avons identifié quatre cas.

ID	Type de problème
4	La licence du jeu de données ne nous permet pas de l'utiliser
6	Le format du jeu de données ne permet pas d'ouvrir le fichier dans des outils très répandus (Excel, Notepad...)
10	Le fichier est mal formé
116	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)

ID	Type de problème	Exemple
4	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

Description du problème et de ses impacts
Ce problème ne concerne pas les données en soit. Ce cas n'est pas si rare et il peut être méconnu des usagers, par exemple dans le cadre d'une grosse organisation où la donnée peut parfois circuler sans information claire sur sa provenance.

Détection
La documentation du jeu de données doit indiquer les conditions d'usage. En son absence, il convient de retracer l'origine de la donnée.

Correction
Il est temps de régulariser sa situation en se tournant vers le producteur des données. Peut-être est-il opportun de regarder si ces données n'existent pas sous d'autres formes, par exemple en open data.

ID	Type de problème	Exemple
6	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

Description du problème et de ses impacts
Ce problème ne s'applique pas seulement aux néophytes. Même un fichier CSV, pourtant recommandé comme format basique pour la circulation des données, peut poser problème à ses utilisateurs.

Détection
Il s'agit d'un contrôle purement manuel consistant à regarder d'une part si le fichier proposé s'ouvre bien et, d'autre part, s'il existe une version adaptée aux logiciels les plus courants.

Correction
Il suffit de proposer aussi un fichier adapté au plus grand nombre, par exemple les formats XLS ou ODS, immédiatement lisibles par les tableurs.

ID	Type de problème	Exemple
10	Le fichier est mal formé	Pour certaines lignes, parfois une colonne manque, ou, le fichier CSV comporte des "virgules" non formatées et empêche l'ouverture correcte du fichier

Description du problème et de ses impacts
Ce problème est rare mais pas si exceptionnel que ça. Il arrive par exemple lorsqu'un jeu de données est exporté automatiquement en différents formats : certains peuvent fonctionner quand d'autres seront mal formés et empêcheront donc tout usage. Autre cas, une modification de la chaîne de production des données peut avoir ce genre d'effet de bord et le producteur ne le remarque pas tout de suite car le process de mise à jour des données est entièrement automatisé.

Détection
Il s'agit d'un contrôle manuel consistant à regarder si le fichier proposé s'ouvre bien. Il est probablement possible d'automatiser ce contrôle en fonction du type de fichier. L'outil csvclean, du paquetage csvkit, permet de contrôler un fichier mal formé en ligne de commande.

Correction
Il convient de revoir la chaîne de production des données.

ID	Type de problème	Exemple
116	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

Description du problème et de ses impacts
Le mode d'accès aux données peut constituer un frein voire une barrière infranchissable à l'usage. Par exemple certaines données ne sont accessibles que via des formulaires et cela rend plus difficile l'extraction (en masse) de l'ensemble des données.

Détection
Là encore, il s'agit d'un contrôle manuel consistant à regarder le mode d'accès au fichier. Selon les modalités d'accès, il est possible d'automatiser ce contrôle. Un robot HTTP peut par exemple vérifier régulièrement la disponibilité d'un fichier : les outils unix curl et wget peuvent être automatisés. Il existe par ailleurs de nombreux services commerciaux dit de "monitoring" de sites web.

Correction
Il convient de revoir le mode d'accès aux données.

Les problèmes liés aux caractéristiques du fichier

Ces problèmes touchent aux deux caractères essentiels d'un fichier : son format et son encodage. Nous en avons retenu 4.

ID	Type de problème
1	Le jeu de données est dans un format "image" ne permettant pas de manipuler les données
2	Le jeu de données est dans un format non spécifiquement adapté aux données : PDF, Word, ODF, epub, HTML, SVG, etc.
5	Le format du jeu de données n'est pas ouvert
8	L'encodage n'est pas en UTF-8 : ce dernier devient la norme de facto et d'autres encodages peuvent engendrer des problèmes

ID	Type de problème	Exemple
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
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

Description du problème et de ses impacts

"On part de loin" s'étonneront certains lecteurs. Mais il ne faut pas oublier que beaucoup de documents sont fournis dans des formats images (ID1) ou des formats non adaptés aux données (ID2). Par méconnaissance de ce qu'est une donnée, certains producteurs ne voient pas la différence. D'autres producteurs y voient une manière de limiter voire d'empêcher la diffusion trop large de leurs données. D'autres encore n'ont pas les moyens de produire autre chose, comme lorsqu'un logiciel métier ne propose que ces formats d'export.
Faut-il le préciser, l'impact est considérable :

- dans l'ID1 il rend le jeu impossible à utiliser à moins d'un long travail de ressaisie ou de reconnaissance de caractères,
- dans l'ID2 il oblige à une préparation des données pour pouvoir être manipulées.

Détection

Ce contrôle amont est manuel et immédiat, il suffit de regarder et/ou d'ouvrir le type de fichier livré. Pour industrialiser cette détection, il est possible d'utiliser le programme unix "file". Ou bien d'utiliser "pdf2txt", du package python "pdfminer".

Correction

On commencera par se tourner vers le producteur des données, peut-être est-il en mesure de fournir les données sous une forme manipulable et spécifiquement adaptée aux données. Dans le cas contraire, outre la ressaisie, la reconnaissance optique de caractères peut être option (ID2) mais les résultats pour reconnaître des tableaux peuvent se révéler très mauvais.
Dans le cas de l'ID2, il est possible de traiter ces fichiers à la main.

ID	Type de problème	Exemple
5	Le format du jeu de données n'est pas ouvert	Le fichier n'est disponible qu'au format .xls ou .xlsx

Description du problème et de ses impacts
Beaucoup de formats propriétaires peuvent être lus, exportés, traités par de nombreux outils. Mais les formats propriétaires oblitèrent néanmoins la pérennité des données. Il arrive que certains formats ne soient lus que par le logiciel qui les a produits : rien ne garantit alors la possibilité future de continuer à exploiter les données.

Détection
Ce contrôle amont est manuel et immédiat, il suffit de regarder et/ou d'ouvrir le type de fichier livré.
Pour industrialiser cette détection, il est possible d'utiliser le programme unix "file".

Correction
On commencera par se tourner vers le producteur des données, peut-être est-il en mesure de fournir les données sous un format ouvert (CSV est recommandé pour les données tabulaires). Il est possible de produire soit-même ce format.

ID	Type de problème	Exemple
8	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

Description du problème et de ses impacts
Version courte : en 2018, il n'y a (presque) plus de raison d'utiliser un autre encodage que UTF-8.

Version longue : L'encodage d'un fichier c'est la norme utilisée pour coder chaque caractère par une suite de 0 et de 1 compréhensible par une machine. L'encodage est le premier facteur de difficulté d'usage : il oblige les réutilisateurs à des opérations de conversion laborieuses ; certains outils ne comprennent pas certains encodages ; etc.
Au début de l'informatique le code dominant était l'ASCII américain. Mais ce dernier ne permettait pas d'encoder les caractères accentués des alphabets latins (français, allemand, etc.) et *a fortiori* les caractères extra-latins. Après de longues années passées à créer et utiliser des encodages "locaux" — comme l'ISO-Latin-1 spécifique au français —, l'UTF-8 a été créé pour coder "l'ensemble des caractères du « répertoire universel de caractères codés »" (source Wikipédia, article UTF-8). Ce dernier est compatible avec l'ASCII et permet d'encoder tous les alphabets extra-latins (grec, japonais, arabe, hébreux…). L'UTF-8 est en passe de devenir le standard de référence universel.

Détection
De nombreux éditeurs de texte signalent l'encodage d'un fichier.
Pour industrialiser cette détection, il est possible d'utiliser le programme unix "file".

Correction
Le mieux est d'agir sur le système d'information à l'origine des données : corriger son encodage de traitement ou d'export est un investissement durable. Dans le cas contraire, il existe de nombreux outils capables de réaliser la conversion. L'utilitaire unix iconv convertit de volumineux fichiers en quelques secondes : `iconv -f ISO-8859-1 -t UTF-8 fichier.csv > converti.csv`

Les problèmes liés à la gouvernance des données

Les problèmes de gouvernance des données ont d'importants impacts sur la qualité des données. Nous proposons 5 points de contrôle.

ID	Type de problème	Exemple
24	Le processus 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
26	La disponibilité de la donnée n'est pas mesurée (temps pendant lequel la donnée est accessible par rapport au temps total souhaité, généralement exprimé en pourcentage)	Le producteur ne sait pas si la qualité de service est de 95% ou 99,99 alors que tel futur usage est critique
28	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
29	La qualité de la donnée n'est pas mesurée	Aucune méthode de contrôle n'est mise en œuvre pour mesurer la qualité des données
117	L'incertitude de la mesure n'est pas connue par le producteur	Le producteur des données ne connaît pas la précision de ses mesures

Description du problème et de ses impacts

Ces 5 items, décrits ici conjointement, permettent de contrôler les principaux points de gouvernance porteurs d'un impact direct sur la qualité des données. Sans process de signalement d'erreur, sans mesure de la disponibilité, sans dispositif de mesure de la qualité (ID28 et 29), sans analyse de l'incertitude des mesures, la qualité n'est pas gouvernée, maîtrisée, suivie. Un usage ponctuel de ces données est déjà problématique. Un usage régulier extrêmement déconseillé : la disponibilité des données peut ne pas être adaptée aux usages ; l'absence de remontée d'erreur empêche toute montée en qualité.

Détection

Si vous êtes réutilisateur des données, vous n'en verrez que la partie immergée : l'absence de documentation de ces points. Si vous produisez ces données, il est plus qu'utile de vérifier si ces aspects sont documentés ou, a minima, rapidement documentables.

Correction

Si vous êtes réutilisateur, nous vous reportons au points de contrôle correspondant à la documentation ci-après.
Si vous êtes le producteur de données ou un service de contrôle, c'est le moment de songer à instaurer quelques process de bonne gouvernance. Il peut s'agir, dans un premier temps, de réunir toutes les parties concourantes à la production pour résoudre et documenter ces aspects.

Les problèmes liés au design ou à la production

La conception et les modalités de production des données engendrent des impacts très divers. Nous avons retenu 9 points de contrôle.
Hors l'homogénéité de l'encodage (ID9), la détection de ces points n'est pas automatisable et peut être assez longue. Par ailleurs, ce sont des problèmes souvent difficiles à corriger car il peut être difficile de revenir sur la conception des données ou leur chaîne de production.

ID	Type de problème	Exemple
9	L'encodage n'est pas homogène	Certaines données sont correctement encodées et d'autres contiennent des caractères ésotériques
19	L'origine de certaines données est une entrée manuelle non contrôlée	Le risque est d'obtenir 25 orthographes de "Saint-André-des-Arts"
20	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 ; 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.
21	L'échantillon est biaisé	Certaines populations sont absentes, sur-représentées ou sous-représentées ; les données subissent une forte variation saisonnière
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
30	Une entité possède plusieurs identifiants	Les associations peuvent avoir un numéro d'association ET un code SIREN unique (problème de design ?)
46	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 de 4 caractères ou plus
47	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

Les problèmes liés à la documentation du jeu de données

La documentation est un vecteur insidieux de non-qualité des données. Nous voyons fréquemment des jeux dont les données sont de très bonne qualité, mais dont la documentation est si pauvre, voire inexistante !, que leur usage en devient problématique. Les problématiques sont extrêmement diverses mais ont presque toutes en commun de posséder un impact très négatif : par exemple quand le format de fichier ou l'encodage n'est pas précisé (ID3 et ID7), quand la documentation est un PDF "image" où il est impossible de chercher (ID115) ou bien encore quand le process de signalement d'erreur n'est pas précisé (ID23).

Nous ne revenons pas ici dans le détail de chaque cas, **l'impact** sur les usages étant invariablement un frein voire une impossibilité d'utiliser les données pour un usage professionnel.

Faute d'une normalisation efficace et répandue de la documentation des données, le **contrôle** de chaque point sera manuel, par une lecture attentive de la documentation existante. Un contrôle approfondi devrait être effectué avant la mise en circulation d'un nouveau jeu de données.

La **correction** de ces problèmes n'est pas compliquée mais demande une participation des différents acteurs de la chaîne de production et d'édition.

ID	Type de problème	Exemple
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
7	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
31	La documentation et les métadonnées sont quasi inexistantes voire absentes	La documentation tient sur 5 lignes alors que le fichier est très complexe
115	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
14	Le process d'acquisition n'est pas connu (il peut être gênant de n'avoir aucun recul critique sur cet aspect)	WikiLeaks
18	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° ?)
23	Le process de signalement d'erreur et d'échange avec le producteur n'est pas explicité	Aucune forme de contact n'est donnée
25	La disponibilité de la donnée n'est pas documentée (temps pendant lequel la donnée est accessible par rapport au temps total	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

	souhaité, généralement exprimé en pourcentage)	devraient en être informés pour savoir si leur usage en est impacté
27	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
32	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 ? * "Résultats des élections à Montréal" : il existe 6 communes appelées Montréal dans le monde...
38	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
39	Manque de métadonnées : la fraicheur des données n'est pas explicité : * le délai entre le réel et la mise en base de la donnée * 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 mois avant d'arriver au réutilisateur
118	La documentation ne précise pas la date de péremption ou d'obsolescence des données	On ne sait pas quand des données deviennent inutile et pourraient donc être effacées ou archivées
120	La volumétrie des données n'est pas précisée	Le poids du fichier n'est pas indiqué ; le nombre de lignes du fichier n'est pas précisé

Les problèmes liés à la documentation des données (contenus, champs)

La documentation des données elles-mêmes ressort de la même problématique que la document du jeu de données. Nous ne détaillons pas ici chaque point de contrôle retenu. Il est très utile de traiter ces points en même temps que ceux retenus pour la documentation du jeu.

ID	Type de problème	Exemple
15	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
16	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. * Un jeu de données contient un champ "Image" en binaire, dont le format n'est pas spécifié.
17	La taille maximale d'un champ n'est pas documentée	On ne sait pas si un code peut dépasser 10 caractères et si certaines valeurs sont donc fausses
33	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.
34	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.
35	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.
36	Manque de métadonnées : le fait que le champ soit un booléen n'est pas spécifié	
37	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"
40	Manque de métadonnées : la langue des textes n'est pas spécifiée	
41	Métadonnées imprécises : le format de date n'est pas spécifié	Format américain ? anglais ? européen ? etc.
42	Métadonnées imprécises : unités non spécifiées	On ne dit pas si colonne "hauteur" est en cm ou dm
43	Métadonnées imprécises : système de coordonnées non spécifié	La documentation n'indique pas si les coordonnées sont en WGS 84, Lambert ou un autre système
44	Métadonnées imprécises : noms de colonnes ambigus	"Emplacement" ne dit rien sur la donnée attendue : une adresse ? "en haut" ? "devant" ? etc.
45	Métadonnées fausses	

Les problèmes liés à l'écosystème de la donnée : standards, normes, codes, etc.

Les données numériques existent depuis plus de 50 ans et un véritable écosystème s'est construit autour d'elles : standards de fichiers, formats de types de données (dates, etc.), référentiels, codes, etc. Cet écosystème participe directement de la qualité des données. Nous avons retenu 9 points de contrôle que nous ne détaillons pas du fait qu'ils partagent les mêmes impacts, les mêmes méthodes de détection et de correction.

S'agissant des **impacts**, de prime abord, ces problèmes ne sont jamais rédhibitoires et ne paraissent pas être si bloquants pour utiliser des données de qualité. En regardant plus précisément, nous voyons bien qu'ils sont nombreux à constituer un gros frein aux usages. Que justifie encore en 2018 de ne pas produire de fichier au format GTFS s'agissant de données de transport ? L'usage de formats complexes ou peu utilisés grève durablement le potentiel d'usage de vos données, aussi qualitative soient-elles.

La **détection** de chaque problème est très simple : un rapide coup d'oeil à la documentation et aux données elles-mêmes conduit à contrôler chaque point — à condition d'avoir quelques connaissances sur les standards mentionnés ; si vous ne les avez pas vous-mêmes, la notoriété et la diffusion de ces standards est telle que vous trouverez facilement des ressources ou des personnes qualifiées pour vous aider.

La **correction** de ces problèmes qualité, très structurants, peut en revanche être longue et difficile en fonction des chaînes de production des données. Pour éviter d'avoir à les traiter ou bien pour veiller à le faire progressivement, nous recommandons que chaque problème soit traité par l'instance ou le process de gouvernance des données.

> OpenDataFrance, par exemple, mentionne explicitement certains de ces standards dans ses *Recommandations pour favoriser l'interopérabilité des données open data* : ISO 8601 pour les dates (ID49 ci-dessous), ISO 639 pour les langues (ID50), ISO 3166 pour les pays (ID49), ISO 4217 pour les monnaies (ID52), WGS84 pour les coordonnées (ID48) — https://frama.link/ODF-reco-interop.

De manière plus générale la gouvernance des données doit s'assurer qu'un jeu de données est ouvert à l'écosystème des données, à travers des formats, des données pivot ou des champs standardisés.

ID	Type de problème	Exemple
121	Le jeu de données ne contient pas de données pivot (données de référence) facilitant le croisement avec d'autres données	Ce fichier relatif aux lycées ne contient pas le code UAI or il existe plusieurs lycées Paul Claudel, rendant les croisements difficiles
11	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
12	Le jeu de données concernant des œuvres n'est pas au format Dublin Core	Le fichier n'est pas au format Dublin Core
13	Le jeu de données utilise une norme peu accessible au plus grand nombre (coût, complexité)	Le jeu de données est au format TRIDENT

48	Les coordonnées ne sont pas au format WGS 84	Les coordonnées sont au format Lambert II nécessitant une conversion des points pour des usages mobiles liés à des GPS grand public
49	Les codes pays ne sont pas au format ISO 3166	L'Allemagne est noté "ALL" alors qu'il existe un code ISO employé internationalement
50	Les codes de langues ne sont pas au format ISO 639	Le français est noté "F" ou "français" en lieu et place de "fr"
51	Les dates ne sont pas au format ISO 8601	La date est notée "01/01/2016"
52	Les monnaies ne sont pas au format ISO 4217	Le franc suisse est noté FS

2. Les problèmes syntaxiques et morpho-syntaxiques

Les problèmes syntaxiques et morpho-syntaxiques sont relatifs à la **manière d'écrire les données** et non leur sens. Ils sont nombreux et extrêmement fréquents. Ils ne sont généralement pas rédhibitoires pris séparément et en faible nombre, mais leur **fréquence et leur addition** peuvent conduire à rendre un jeu de données presque inexploitable : ils peuvent avoir un fort **impact** sur l'analyse et la manipulation des données, comme les tris, le dédoublonnage, les statistiques voire le simple comptage de certaines choses.

Erreurs et incohérences syntaxiques

Les problèmes de syntaxe apparaissent dans deux cas :

- lorsqu'il existe une syntaxe pour une certaine donnée : par exemple, pour les nombres, on parle alors d'**erreurs de syntaxe ;**
- ou bien lorsqu'il n'existe pas de syntaxe claire prévue pour une donnée et que plusieurs syntaxe coexistent, créant alors des **incohérences syntaxiques.**

Les erreurs de syntaxe sont faciles à détecter et parfois à résoudre :

- manuellement, il s'agira de rechercher des chaînes de caractère qui peuvent poser problèmes ;
- semi-automatiquement, il est possible d'obtenir une analyse très fine et rapide à partir de regex (voir l'annexe en fin de document).

Les secondes, en l'absence de syntaxe existante, exigent un travail d'investigation qui peut être très important. Elles donnent l'occasion d'évaluer si des actions préventives doivent être menées en amont, typiquement le fait de spécifier une syntaxe.

Nous avons identifié 11 types de problèmes, 8 cas d'erreurs et 3 cas d'incohérences fréquents. On pourrait identifier de très nombreux autres cas, nous avons cru bon de ne retenir que des cas fréquents et faciles à détecter et traiter.

ID	Type de problème	Exemple
53	Erreurs syntaxiques : espace(s) au début ou à la fin du champ	" Pierre" au lieu de "Pierre"
54	Erreurs syntaxiques : bug syntaxique dans les strates du SI : le cas de l'apostrophe	"N\'Diaye" à la place de N'Diaye
55	Erreurs syntaxiques : syntaxe des numéros ou nombres en tous genres	"45€" au lieu de "45"; "1,000,000" au lieu de "1000000"
56	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
57	Erreurs syntaxiques : les sigles et abréviations ne sont pas homogènes	"SNCF", "S. N. C. F.", "S.C.N.F." ? "Boul" ou "Boul." ou "Bld" ?
58	Erreurs syntaxiques : booléen	"V" au lieu de "1" selon la spécification du booléen

59	Erreurs syntaxiques : email, url	laurent.dupont@wanadoo@fr
60	Erreur syntaxique sur la date	2016/09/30 au lieu de 2016-09-30 attendu
61	Incohérences syntaxiques : syntaxe des noms propres	"de La Tour" ou "La Tour (de)" ?
62	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"
63	Incohérences syntaxiques : l'usage du pluriel ou du singulier	On trouve parfois la catégorie “boulanger” et parfois “boulangers”

Les erreurs syntaxiques [75%]

Toute **chaîne de caractère** stockée ne devrait pas contenir d'espace en début ou en fin chaîne.

Numéros ou nombres en tous genres :

- certains utilisateurs ou opérateurs de saisie utilisent régulièrement la lettre "o" en lieu et place du chiffre zéro
- dans certains jeux de données on voit l'unité qui caractérise la valeur (par exemple 45€, 35 km ou bien encore 2,5 Wh) : cette pratique est à éviter
- l'usage du point ou de la virgule est regarder de près
- le séparateur de milliers, qu'il s'agisse d'un espace, d'un point ou d'une autre forme, est à proscrire au moment de la saisie et en tant que donnée stockée ; à l'inverse une interface peut l'ajouter automatique à des fins de lecture uniquement (en France l'usage est plutôt d'utiliser un espace comme séparateur)

Code postaux : 7100 n'est pas un code postal valide. Un code postal est toujours à 5 chiffres, il faut donc écrire 07100.

De très nombreux **codes** ont une syntaxe bien précise qu'il est utile de connaître. On regardera par exemple les codes suivants :

- code commune INSEE
- numéro de sécurité sociale
- SIRET
- code APE
- etc.

Les accents sur les capitales : si je liste les relations des personnages de Game of Thrones : la relation "TUE" signifie-t-elle "tue" ou "tué" ? Hélas, des générations d'enseignants déformés par les machines à écrire aux capitales non accentuées ont appris à des générations d'écoliers "*qu'on ne met pas d'accent sur les capitales*". C'est une grossière erreur ….

Emails : les emails suivent des règles strictes[4] ; idéalement un email est contrôlé au moment de sa saisie mais cela n'est pas toujours fait ; voici les erreurs communes :

- espace dans une adresse
- accent dans une adresse ; c'est théoriquement possible puisque les accents arrivent dans les noms de domaine ; la pratique est rare mais elle existe[5]
- capitales dans une adresse ; elles n'affectent pas le bon fonctionnement de l'adresse email mais posent des problèmes de tri, de faux pseudo-doublons, etc.
- absence d'arobase
- arobase placée au tout début
- la présence de deux points à la suite : "jean..chaple!@example.com"

Une **URL** ……..

[4] La RFC 6530 : https://tools.ietf.org/html/rfc6530 ; Wikipedia en donne une glose accessible, en français : https://fr.wikipedia.org/wiki/Adresse_%C3%A9lectronique

[5] Le service Gmail reconnaît de telles adresses depuis 2012.

Les incohérences syntaxiques [75%]

(DU PONT, Du Pont, Pont (Du), 01.40.02.04.55, 01-40-02-04-55, 01 40 02 04 55,, etc.)
Les incohérences ne sont pas des erreurs à proprement parler, d'autant plus lorsque la syntaxe et la sémantique des données n'a pas été normalisée et/ou que l'outil de saisie ne les contrôle pas.
Mais les incohérences ont de nombreux impacts négatifs sur la qualité actuelle et les usages futurs :

- tri, recherche
- exemplarité

Les **noms propres** doivent faire l'objet d'une attention particulière :

- St-Mandé, St Mandé, Saint-Mandé, Saint Mandé, ST-MANDE, ST MANDE, SAINT-MANDE, SAINT MANDE, ST-MANDÉ, ST MANDÉ, SAINT-MANDÉ ou SAINT MANDÉ ? Soit 12 façons de l'écrire !
- Provence-Alpes-Côte d'Azur, Provence-Alpes-Côte-d'Azur ou PACA ?

Pour les **noms communs l'usage du pluriel ou du singulier** peut poser problème. Par exemple, pour qualifier une cargaison, devra-t-on saisir "Tomate" ou "Tomates" ?
Certains recommandent l'usage du singulier seul même si certains cas ne s'y prêtent pas ; par exemple "archives" ou "chaussures" sont généralement utilisés au pluriel.

Les **sigles et acronymes** doivent faire l'objet d'une attention particulière : SNCF vs S. N. C. F., ONU vs Onu vs O. N. U. vs O.N.U vs O.N.U. : il y a tant de manières d'écrire un sigle ou un acronyme !

Les dates peuvent aussi s'écrire de manières très différentes

- 20/06/1969
- 20-06-1969
- 20-06-69
- etc.
- 20 juin 69
- 19690620
- 06/20/1969 (au format britannique, le mois étant placé devant)

Les **numéros de téléphone** peuvent faire l'objet de syntaxes différentes.

- 06 16 72 72 38
- 0616727238
- 06-16-72-72-38
- 06 1672 7238
- 06 16 7272 38
- 06 166 272 38
- +33 6 16 72 72 38

[...]

Les problèmes morpho-syntaxiques : l'usage de la forme pour exprimer le fond

Ces problèmes sont à la frontière entre la syntaxe et la sémantique. Le cas typique est l'usage de la couleur, dans un tableau Excel par exemple, pour exprimer des valeurs. C'est une erreur plus fréquente que l'on croit. Qui n'a jamais utilisé de rouge pour signaler une donnée particulièrement sensible ou bien du vert pour signaler un chantier terminé ? Que se passera-t-il lorsque ces données seront exportées, par exemple au format CSV ?

FRÉQUENTATION DE 2006 à 2010 (Sour

N°REF	ETABLISSEMENTS	VILLE	TOTAL 06
6816401	MUSEE DU VIGNOBLE ET DES VINS D'ALSACE	KIENTZHEIM	3 502
6822401	MUSEE DES BEAUX-ARTS	MULHOUSE	13 085
6822402	MUSEE DE L'IMPRESSION SUR ETOFFES	MULHOUSE	31 437
6822403	MUSEE NATIONAL DE L'AUTOMOBILE - Musée national - Collection Schlumpf	MULHOUSE	187 000
6822404	MUSEE ELECTROPOLIS	MULHOUSE	34 323
6822405	MUSEE FRANCAIS DU CHEMIN DE FER - CITE DU TRAIN	MULHOUSE	100 537
6822406	MUSEE DU SAPEUR POMPIER	MULHOUSE	Fe
6822408	MUSEE HISTORIQUE	MULHOUSE	35 018
6822409	MUSEE MINERALOGIQUE DE LA SOCIETE INDUSTRIELLE DE MULHOUSE	MULHOUSE	130
	8	MULHOUSE	401 538
6823101	MUSEE VAUBAN	NEUF-BRISACH	2 165
6824801	MAISON DU SUNDGAU - MUSEE PAYSAN	OLTINGUE	1 566
6827702	MUSEE DU DOLDER	RIQUEWIHR	4 620
6827801	MUSEE DU PAPIER PEINT	RIXHEIM	9 235
6833401	MUSEE DES AMIS DE THANN	THANN	7 925
6834301	ECOMUSEE D'ALSACE	UNGERSHEIM	
48	MUSEES		1 325 982
		nb réponses par année	46

Données confidentielles - Contacter le chef d'établissement

Appellation en 2007

Appellation en 2008

Chiffres de fréquentation de la collection permanente. Ouverture du musée uniquement sur rende

Gratuité exceptionnelle durant toute l'année 2007 pour le centenaire du musée. Musée fermé du 2

Ouverture du musée au printemps 2011

Nous avons regroupés ces problèmes sous la forme de 2 points de contrôle (ci-dessous) que nous ne détaillons pas séparément tant ils sont proches.

L'usage de la forme pour le fond provoque trois catégories d'**impacts** :

- les données deviennent **difficile à manipuler** : avez-vous déjà essayé d'additionner des cases rouges sur Excel ?
- les données sont souvent plus **mal documentées**, car justement elles sont d'un type différent de toutes les autres données
- ces données sont **fragiles** : par exemple elles résistent très mal à un export car le logiciel qui les a produit est souvent le seul à les interpréter ; beaucoup de format de données ne possèdent pas même la possibilité d'une mise en forme, comme le format CSV par exemple.

Leur **détection** manuelle est souvent très simple :

- les problèmes apparaissent visuellement dès l'ouverture (couleurs, fusions, etc.)
- l'export des données produit un fichier corrompu (du fait des fusions par exemple)

Leur détection automatique, en revanche, n'est pas simple car elle dépend du logiciel qui a produit les données. Un fichier am formé (ID10 ci-dessus), peut être le signe de fusion de cellules.

Leur **correction** est souvent assez simple mais forcément manuelle et elle peut se révéler longue.

ID	Type de problème	Exemple
77	Exprimer une donnée à travers un code difficile à manipuler	Mise en forme pour exprimer une donnée : couleur, gras, etc.
78	autres ?	Certains fichiers possèdent des cellules fusionnées ; des données sont ajoutées sous forme de commentaires ; etc.

3. Les problèmes sémantiques [10%]

Dans une application bien pensée, on ne devrait pas avoir d'erreur sémantique. Par exemple le genre peut se définir à l'avance selon les codes H ou F pour "homme" ou "femme" ; dans ce cas l'application ne devrait pas permettre d'entrer "M" pour masculin. Mais il arrive fréquemment que des données soient agrégées …….

L'inversion de deux valeurs fait aussi partie des erreurs sémantiques courantes, par exemple :

- l'inversion de deux coordonnées GPS
- l'inversion du prénom et du nom

ID	Type de problème	Exemple
64	Plusieurs termes sont utilisés pour un même sens	Parfois on lit "Daesh", parfois "Isis" et parfois "EI" ; ou bien "agent" ou "commercial" ; etc.
65	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 "États-Unis" et pas "USA" qui est un terme anglais
!66	Certains termes, valeurs utilisées sont vieillis, inusités, cryptiques ou incompréhensibles	
67	Les abréviations ou sigles ne sont pas explicités	Wikipédia fournit des listes de très nombreux sigles : https://fr.wikipedia.org/wiki/Sigle
68	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
69	Inversion dans un couple de données	"Dupont Jean" au lieu de "Jean Dupont'
70	L'absence de lettres accentuées peut poser des problèmes de sens	"JUPE TUE LA FRANCE GAGNE"
71	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	Erreur de système de coordonnées	Coordonnées en Lambert II au lieu de WGS 84 spécifié dans les métadonnées
73	Les coordonnées géographiques sont données en degrés, minutes, secondes et non en degrés décimaux, ce qui complique leur manipulation	23°56'33" ou bien 23°56'33"E en lieu et place de la forme décimale 23,9756
74	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)
75	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.
76	Liste de réponses fermée mal conçue : présence de la réponse "Autre" ou "Divers" très fréquente	"Quel est votre ville favorite : Marseille, Paris, Autre"

Les erreurs de contenu/validité des contenus [10%]

Ce sont probablement les plus nombreuses et les plus difficiles à détecter et corriger.

79	Aberration	* 197 ans (pour l'âge d'une personne) * Général de Gaulle comme personne participant à un sondage
80	Doute très raisonnable, valeurs inexplicables	20 participants de plus de 110 ans
81	Certaines valeurs sont suspectes : 0000 ou xxxxxxxxxxxx (à compléter)	-
82	Certaines valeurs sont suspectes : suites de chiffres comme 9999 ou 12345	Des suites de 9999 ; nombreuses valeurs "12345" (détailler)
83	Certaines valeurs sont suspectes : il existe des dates en 1900, 1904, 1969, 1970	-
84	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	La source n'est pas crédible (incompétent, juge et partie, etc.)	15000 manifestants selon les organisateurs
86	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.)

Les données périmées [75%]

Les problèmes et leurs impacts

Quelques exemples pris pour leur diversité
Beaucoup de valeurs liées au temps peuvent devenir fausses sans information de contexte : le 1er au classement de la ligue 1 de Football peut changer toutes les semaines (!).

Les cas les plus classiques :

- saisie fausse
    - le sexe
    - l'âge
    - l'adresse postale
    - l'adresse mail
    - les téléphones
    - les membres du foyer (de nombreuses pratiques commerciales sont liées au foyer)
- les données suspectes
    - dates de naissance au premier janvier

Est-il possible d'évaluer l'importance de ces problèmes qualité : nombre, coût, etc. ?

- Hyper compliqué à faire.
- Le curatif est plus facile à faire que le préventif.
    - Une petite équipe au niveau du décisionnel est là pour conduire les démarches curatives et pour juger à la louche s'il faut lancer des chantiers préventifs
    - Souvent le préventif est long, compliqué et impossible par les gens qui sont en bout de chaîne, les usagers des données
- Certains "cas" ont un gros impact interne :
    - le client peut suivre les données d'un homonyme
    - le couple se sépare et M. ou Mme contenu de voir les contrats du conjoint
        - Ces cas là font l'objet d'un traitement de crise : curatif ou préventif
        - parfois les cas de crise "passent" et ne sont pas traités

NPAI (maintenant PND (pli non distribué)) très faible.

Certaines sont faciles à identifier, par exemple :

- bibi.fricotin@caramail.fr est une adresse qui n'existe plus, le fournisseur ayant fermé depuis plusieurs années
- une date de naissance de plus de 130 ans pour une personne vivante, est probablement erronée
    - impact commercial : il y a un abattement commercial pour les personnes de plus de 90 ans

D'autres sont devinables et contrôlables : un prix qui a plus de 6 mois a toutes les chances d'être faux.

Exemples :

- nombre d'enfants qui n'est pas correct dans un foyer (ils ont pris leur autonomie)
- les dates ne sont pas toujours contraintes car il faut arbitrer entre l'usage et la qualité
    - comment améliorer sans trop contraindre l'usage
        - champs déroulant

        - champs contraint
        - contrôles batch a posteriori
        - icône pour le sexe
        - on fait la chasse au clic dans les organisations car cela perturbe le discours commercial
- les déclarations de naissance ne sont pas pris en compte dans le système
    - source : rafraîchissement de l'information par les conseillers
    - parfois on a l'information dans 2 systèmes différents et ces derniers ne sont pas interconnectés
- il y a des réticences à actualiser des données contractuelles en automatique
    - exemple un changement de code postal peut entraîner un changement de contrat et donc une mécanique d'envoi de nouveau contrat
    - au tout début on avait un seul système, puis deux avec un

Les données trompeuses : robots ou données fausses des utilisateurs [90%]

Description du problème qualité et de ses impacts

Les robots peuvent laisser des adresses postales, des adresses emails, des numéros de téléphones.

Exemples

“On a des robots qui viennent nous visiter pour obtenir des tarifs types”.
5000 personnes arrivent d'un coup de type titi01@wanadoo.fr, titi02@wanadoo.fr ou titi@wanadoo01.fr et qui testent les tarifs pour de nombreuses villes.
On voit parfois arriver en pleine nuit 5000 devis sur des profils similaires probablement pour voir le positionnement tarifaire.

Impacts

Cela pollue les bases avec des faux comptes. Ça fausse complètement le travail d’analyse statistique. Après chaque devis, il y a une équipe qui fait des relances dans les jours qui suivent pour transformer le prospect en client. Pour chaque devis il y a un coût généré qui n’est pas négligeable. Cela a un impact considérable sur les taux de transformation : c’est un problème pour l’entreprise dans son ensemble mais aussi pour les opérationnels dont les objectifs de production sont faussés (avec de possibles conséquences salariales ou d’évolution). Le travail de nettoyage de ces informations perturbent les différents indicateurs et exige un temps de traitement qui doit être déployés rapidement.

Détection

Repérer les fluctuations d’activités importantes, repérer les chiffres dans le mail, les séries d’un mail à l’autre et d’un téléphone à l’autre, repérer les mêmes prénoms+nom en masse.
Il y a un vrai travail d’analyse à la fois visuel et statistique.

1. Au début, on identifie visuellement les suites puis des contrôles automatiques des suites ont été mis en place sur 1 à n caractères (1, 2, 3, 4, ou a, b, c, d, etc.).
2. Les prénom-nom en masse : exemple 500 dossiers arrivent avec “Titi Toto”
3. Les extensions possibles d’adresses mails (.fr, .eu, etc.) et noms de domaines (wanadoo.fr, free.fr, etc.)
4. Faire une commande “whois” sur les noms de domaine pour s’assurer que les noms de domaine existent.
5. Identifier les emails temporaires (exemple : http://www.yopmail.com) ; les personnes peuvent les utiliser pour de très bonnes raisons, il peut donc être utile de les “tiper” et non de les supprimer à court terme. Ils facilitent les nettoyages en lot une fois les durées expirées.
6. Demander aux serveurs de mails qui le permettent si l’utilisateur existe.
7. Détecter les noms de famille fantaisistes tels que Bob L’EPONGE. S’appuyer sur une liste de noms de familles possibles et leurs probabilités de récurrences : comparer avec une base de noms de famille : 50 dossiers “Nepote” le même jour sont impossibles (nom de famille peu répandu).
8. Détecter les emails mal formés : attention, ils peuvent être des emails erronés ou générés par des robots.
9. Quelque chose sur le rythme de remplissage des formulaires ?
10. L’adresse IP peut être un indicateur pour identifier des pratiques indésirables. Il n’est pas possible d’avoir plus de 20 demandes avec la même adresse IP par jour. L’adresse IP utile pour regrouper les doublons.
11. Augmentation de l’activité inexpliquées (absence de campagne marketing...)

Solutions mises en oeuvre

On détecte et on qualifie

1. On les “type” (on met une marque informatique) : robots, faux prospects, doublons et vrais prospects (ceux dont on pense qui le sont).
2. En fonction du contexte :
    a. On les exclut selon le contexte : au plus tôt pour l’opérationnel, mais pour des raisons statistiques parfois on les conserve.
    b. On les laisse comme ça, ils sont purgés au bout d’un certains temps sans activité

4. Le manque de données

Le manque de données est une des grandes catégories de la qualité. On retrouve ici l'idée que la qualité de certaines données peut être adaptée à un usage, mais que le manque de certaines données rend d'autres usages impossibles.

Nous avons identifié 4 grands cas du manque de données, détaillés dans les pages suivantes.

La donnée est le résultat d'un calcul dont on n'a pas les données de départ

100	La donnée est le résultat d'un calcul dont on n'a pas les données de départ

Les trous

101	Les trous : manque des "enregistrements" : des données dont vous connaissez l'existence sont manquantes
102	Les trous : manque des "enregistrements" : le tableau possède 65536 lignes
103	Les trous : les données d'un champs sont tronquées
104	Valeurs vides dans certains champs

Les problèmes de granularité

105	La granularité n'est pas suffisante
106	Le fuseau horaire n'est pas précisé dans un contexte de données réparties sur des fuseaux horaires différents
107	L'insuffisance en matière de fréquence
108	L'insuffisance en matière de maillage
109	L'insuffisance en matière de fraîcheur

Les données horaires sans fuseau

106	Le fuseau horaire n'est pas précisé dans un contexte de données réparties sur des fuseaux horaires différents

La donnée est le résultat d'un calcul dont on n'a pas les données de départ

ID	Type de problème	Exemple
100	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.

Description du problème qualité et de ses impacts

Ce problème se rencontre lorsque le jeu de données contient un chiffre fruit d'un calcul dont on ne possède pas les données brutes de départ : typiquement un pourcentage, un rapport, une densité, un indice, etc. Par exemple, lors du résultat d'une élection, le fait d'avoir un résultat par commune en % ne permet pas de mesurer sur une autre granularité car la moyenne des pourcentages de différentes communes est différente d'un pourcentage global des communes concernées. Il faut les données brutes de départ.

L'âge, est un autre exemple courant s'il n'est pas contextualisé. Dire que tel individu a tel âge sans indiquer la date de collecte de cette donnée a toute les chances de n'avoir aucun sens. A tout le moins, cette données sera peu exploitable pour d'autres usages.

Même si l'âge est un peu contextualisé -- 18 ans en 2016 --, le manque de précision peut poser de nombreux problèmes dans d'autres contextes d'usages : depuis quand est-il majeur ? A-t-il 18 ans révolus en 2016, les a-t-il eu ou bien va-t-il les avoir pendant cette année ?

La donnée brute est-elle difficile à lire ?

Ces cas sont nombreux : la date de naissance est d'une lecture rapide mais nous pousse à un calculer l'âge mentalement, plus parlant ; des centaines de milliers de votants est difficile à comprendre si on ne les exprime pas en rapport avec le nombre de votants ; etc.
Pour faciliter le travail des usagers, les producteurs sont donc tentés de publier une donnée calculée. Est-ce leur rôle ? Est-ce une raison pour masquer les données brutes ?

La théorie voudrait qu'un producteur de données ne fournisse que les données brutes sans aucun calcul, laissant le soin au réutilisateur d'effectuer les calculs dont il a besoin. De notre point de vue, il existe une position médiane entre "que des données brutes" et des données calculées sans données brutes : au cas par cas il est possible d'offrir tout à la fois les données brutes et les données calculées, certe avec **un peu de redondance**, mais une redondance arbitrée entre confort de l'usager et surcroît de données.

Détection

La détection de ce problème est essentiellement visuelle. On effectuera un contrôle minutieux de la documentation des données pour identifier tous les ratios, pourcentages,

Correction

S'il manque les données brutes, il est probable qu'elles ne soient pas perdues, selon la logique que le calcul n'est probablement qu'une projection des besoins des réutilisateurs par le producteur. Ce dernier les conserve probablement pour ses propres usages. On tâchera donc de le contacter pour connaître l'opportunité de les voir intégrées dans les données.

Les trous

Description du problème qualité et de ses impacts

Les occasions d'observer des trous de données sont nombreuses et variées.
L'**impact** sur les usages peut être dévastateur en amputant une population ou bien faussant les résultats par un champ ou une table tronquée. Ces impacts peuvent être d'autant plus insidieux qu'ils peuvent être relativement discrets et ne se révéler que dans un temps long -- par exemple plusieurs mois après une migration apparemment réussie, conduisant à une perte irrémédiable de données.
Nous avons retenu 3 problèmes.

ID	Type de problème	Exemple
101	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	Les trous : manque des "enregistrements" : le tableau possède 65536 lignes	-
103	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	Valeurs vides dans certains champs	-

Détection

Un contrôle visuel doit généralement permettre de se rendre compte de ce type d'erreur ; un tri alphabétique des colonnes peut grandement aider.
Plusieurs méthodes de contrôles sont simples et peuvent être semi-automatisées :

- alerter automatiquement d'un fichier ou d'une table qui flirte avec 65536 lignes
- alerter automatiquement d'un pourcentage élevé d'enregistrements s'arrêtant pile à 255 caractères (ou bien à X caractères, X représentant la dimension maximale du champ)
- rechercher des erreurs fréquentes comme "l'avenue du Général de Gau"
- produire un rapport statistique automatique calculant le nombre de valeurs uniques par champ

D'autres méthodes sont plus longues ou plus subtiles :

- comparer les valeurs de certains champ avec un dictionnaire ou un référentiel connu
- ...

Correction

La solution est souvent à chercher en amont du problème : sensibiliser les collecteurs, rendre un champ obligatoire, revoir une chaîne de production qui tronque des fichiers, etc.

Certaines méthodes permettent néanmoins de compléter des données a posteriori. Le genre peut par exemple être déduit du prénom. Un champ tronqué, déduit d'un référentiel comme par exemple la fin d'une adresse trouvée par le référentiel adresse d'une commune.

La granularité de l'information et ses avatars : précision, fréquence, maillage, fraîcheur

La granularité est un des problèmes qualité “en creux”, qui n’apparaissent que lors de besoins nouveaux. De fait, une granularité médiocre peut oblitérer de nombreux usages ou bien donner une vision complètement erronée d’un phénomène : il est facile de comprendre par exemple que la mesure du bruit en un point de Paris ne pourra se généraliser. Nous avons retenu ce problème sous sa forme générale (ID105) et sous 3 formes particulièrement fréquentes.

ID	Type de problème	Exemple
105	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
107	L’insuffisance en matière de fréquence	L'état de feux tricolores classiques est donné tous les jours à minuit
108	L’insuffisance en matière de maillage	La pollution dans Paris est mesurée avec un seul capteur
109	L’insuffisance en matière de fraîcheur	Les chiffres du recensement de cette espèce date de 1976

Détection
La détection de ce problème est essentiellement visuelle. On effectuera un contrôle minutieux de la documentation des données pour identifier les catégories et évaluer leur échelle. On regardera les données avec attention, par exemple sous la forme de tris de colonnes : des tris de dates par exemple (pour le cas de la fréquence et de la fraicheur), des tris de catégories pour évaluer la granularité en général. La mise en carte des données est intéressante observer la granularité spatiale.

Correction
Pour traiter ce problème, on réfléchira à l'effort supplémentaire nécessaire pour affiner la granularité. Il peut être très coûteux d’affiner un maillage ou la précision d’une mesure. La fréquence et la fraicheur peuvent en revanche être peu coûteuses si les capteurs sont permanents.

Données horaires sans fuseau

ID	Type de problème	Exemple
106	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

Dans un pays comme la France qui ne connaît qu'un fuseau dans sa version métropolitaine, cette question des fuseaux paraît peu importante. Notre "implantation" planétaire (outre-mer), mais aussi notre marché économique peut rapidement poser des problèmes de traitement et d'analyse lorsque l'information du fuseau manque. Cette information est pourtant très bien standardisée — avec l'ISO 8601 — et certaines professions la manipulent quotidiennement en grande quantité (transports aérien, etc.). Mieux, cette norme ISO est la même que celle qui convient aux dates.

Détection
S'il existe des données horaires, on regardera les métadonnées et les données :

- afin d'apprécier si le fuseau est présent ou non
- afin de voir si un fuseau par défaut est spécifié
- afin de voir si le fuseau peut se déduire sans ambiguïté du contexte de production des données

Correction
En amont, l'information du fuseau est probablement facile à ajouter car de nombreux outils et langages de programmation intègrent l'information dans leur "timestamps". Pour un format compact, ISO 8601 suggère l'emploi de l'heure GMT, sous la forme suivante : 15:49:01Z.
En aval, dans bien des cas il est facile d'ajouter cette information si on est sûr du fuseau. À défaut, il sera utile de spécifier le fuseau par défaut dans la documentation.

Données manquantes ou absence de données ?

Il existe une subtile distinction entre données manquantes et l'absence de données. Dans le premier cas, les **données manquantes**, le terme manquant sous-entend que la donnée existe mais qu'elle manque dans le tableau de données. Des données peuvent manquer pour tout un tas de raisons :

- manque de temps ou paresse du collecteur, typique des systèmes de collectes collaboratifs (OpenStreetMap, Open Food Facts, etc.)
- erreur de transmission entre deux systèmes
- effacement quelconque dans la chaîne de production
- etc.

Ces données peuvent être retrouvées ou complétées a posteriori. Un système de contrôle peut utilement les détecter et suggérer qu'elles soient complétées. Une absence de valeur peut aisément les caractériser, la case vide pouvant implicitement signifier qu'elle attend d'être remplie.

L'**absence de données** peut à l'inverse signifier que la donnée n'existe pas et qu'elle ne saurait être complétée a posteriori. Typiquement dans des tableaux comparatifs de relevés de températures, ces dernières peuvent démarrer à des années différentes selon les lieux. L'absence de valeur peut implicitement signifier l'absence de ces données, mais il peut être utile de **déclarer clairement et explicitement l'absence d'une donnée**, qui ne pourra être complétée a posteriori, puisque la mesure n'existe pas ou n'a pas de sens. Différentes valeurs peuvent convenir du moment que chaque

convention pour désigner l'absence de données, soit bien explicité dans la documentation du jeu de données.

Mathieu Morey signale par exemple que pandas (bibliothèque python) a une liste de valeurs nulles par défaut : '', '#N/A', '#N/A N/A', '#NA', '-1.#IND', '-1.#QNAN', '-NaN', '-nan', '1.#IND', '1.#QNAN', '<NA>', 'N/A', 'NA', 'NULL', 'NaN', 'n/a', 'nan', 'null'.

Dans de nombreux ouvrages statistiques, le trait d'union (-) est souvent employé et présente l'avantage d'être rapide à saisir. Mais ce dernier possède aussi l'inconvénient de pouvoir être mal interprété comme le symbole mathématique "moins" par certains outils. "null" semble avoir la faveur de plusieurs bibliothèques ou langages répandus comme Pandas, R, etc.

À lire ailleurs : Missing Data, de Joe Celko.

La distinction "manquante" ou "absente' peut avoir des conséquences très concrètes. Dans le calcul d'un score, par exemple, le système pourra distinguer :

- le cas où la donnée est explicitement absente (valeur nulle), le calcul du score étant donc complet
- le cas où la donnée est manquante, signifiant que le calcul du score peut être biaisé par ce manque, le système pouvant alors fournir un avertissement

5. La surabondance de données [70%]

La surabondance de données se trouve lorsque des données comportent des informations inutiles voire contre-productives. Le cas le plus typique sont les **doublons**. Mais la surabondance se cache aussi sous d'autres formes : le calcul, la trop grande précision en matière de taille, de fréquence, etc. Ce problème est plus important qu'il n'y paraît et ses impacts sont non négligeables : l'affichage de données trop abondantes peut **allonger les temps de traitements** ; trop de données peut jouer sur la **charge cognitive des utilisateurs**. Trop d'indicateurs rendent le travail d'analyse parfois difficile — comme des données à 6 chiffres alors que 2 suffirait.

Existe-t-il des données trop précises à l'heure où le stockage coûte si peu cher ? Il semble que ce chapitre nous en donne des réponses intéressantes.

ID	Type de problème	Exemple
110	Doublons	* Jean Martin;21/12/1956;Lunéville;noscontacts * Jean Martin;21/12/1956;Lunéville;listesuppl.
111	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	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	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.
114	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
119	La durée de conservation des données dépasse une certaine date de péremption ou d'obsolescence (date légale ou date d'usage opérationnel)	Un prospect de plus de 5 ans a des chances de ne plus avoir aucun intérêt opérationnel

Les doublons [70%]

ID	Type de problème	Exemple
110	Il existe des doublons dans la table : doublons stricts, doublons de ligne ou doublons d'objet	* Jean Martin;21/12/1956;Lunéville;noscontacts * Jean Martin;21/12/1956;Lunéville;listesuppl.

Description du problème qualité et de ses impacts

Encadré 1 : la gestion des doublons de personne à la MAIF : un exemple d'amélioration de la qualité de données en entreprise (par Hugues Roumezin et Jean-Michel Barribaud)

Problématique : le système d'information permet à une même personne d'être identifiée plusieurs fois à plusieurs titres, par exemple :

- titulaire de contrats
- demandeur d'un devis sur le site internet sans s'être identifié au préalable

La présence de personnes identifiées plusieurs fois ont trois types de conséquences :

1. les suivis sont faux (nombre de prospects "gonflé", taux de transformation incorrect, ...)
2. une même personne peut être cible d'une même action commerciale 2 fois ou plus
3. lors d'un appel ou d'une visite, le conseiller aura du mal à identifier la personne parmi plusieurs et prendra plus de temps, sans compter l'image donnée (euh... j'ai 3 Nicolas Sarkozy à l'écran, c'est vous qui habitiez palais de l'Elysée ?)

2 solutions ont été mise en place :
1) Recherche de pré-existence d'une personne : il s'agit d'une aide pour les conseillers du réseau de distribution qui permet une recherche optimisée de la Personne au moment du contact avec celle-ci, via la saisie dans l'IHM relation des données suivantes : Nom, Prénom, Code Postal, Commune, Date de Naissance, Pays (sachant qu'il faut saisir à minima les 2 premières lettres du Nom).
2) Repérage et Fusion des doublons : il s'agit d'un traitement batch qui rapproche des données suivant un score et qui ensuite fusionne les Personnes pour ne garder qu'un seul enregistrement en fonction d'une "matrice de fusion" qui détermine selon des critères la Personne Maître ainsi que les attributs qui lui sont liés et qui doivent être conservés.

Nous avons choisi de mettre en place le produit Informatica Data Quality (IDQ) qui répondait à notre besoin mais qui nous permet également de réaliser le profilage de données ou encore dans le futur des opérations de normalisation / standardisation des adresses postales.

Quelques chiffres

- 3 millions de clients
- 11 millions de personnes (clients + prospects + clients filiales + bénéficiaires + ...)
- Grâce à la mise en place de ces solutions le taux de doublons Personnes Physiques a baissé de 3,5 points sur 4 ans.

Une précision inutile voire contre-productive [70%]

ID	Type de problème	Exemple
111	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.

Ce cas précis est délicat. En temps normal, pour toutes les raisons précisées plus haut, il n'est pas utile d'ajouter une colonne dont la valeur est obtenue par calcul. Les pourcentages sont un exemple typique : d'un certain point de vu, on sent que le producteur des données cherche à éditorialiser un peu ses données pour en faciliter la prise en main. Est-ce son rôle ? Ne détourne-t-il pas le réutilisateur des données de son travail d'analyse et de découverte en lui offrant des choses qu'il n'a pas demandées, qu'il n'attendait pas forcément, et qu'il peut de toutes façons calculer lui-même ?

Ce cas n'est toutefois pas à trancher radicalement. Il existe de bonnes raisons pour indiquer, par exemple, l'âge d'une personne : par exemple un mode de calcul commun d'un âge. Pour des calculs de grille tarifaire c'est basé sur l'âge réel mais pour du décisionnel on est pas obligé d'avoir l'âge précis, on va raisonner par tranche.
Selon l'usage, un peu de surabondance de données est utile. Ajouter des données permet aussi de **soulager les machines pour des calculs complexes**. Quand il faut également normaliser l'usage d'une donnée, par exemple sur de la consommation de produit : une fois qu'on a catégorisé des clients, on ne recalcule pas leur catégorie à chaque fois.

Le cas des données de référence
Les données de référence forment un bon exemple de redondance utile. L'INSEE (institut national de la statistique) se produit par exemple chaque année un Code officiel où chaque commune est dotée d'un code de référence. Ce dernier est réutilisé comme identifiant unique dans des milliers voire des dizaines de bases de données. Mais sa base de référence n'en donne qu'une version éclatée : il faut ajouter le code département (01) au code de la commune (001) pour obtenir la version complète du code (01001). Ce n'est pas grand chose mais cela signifie un temps machine et humain considérable à calculer le code complet qui aurait pu l'être une seule fois en amont.
Il faut aussi considérer tous les utilisateurs non-spécialiste qui ne savent comment concaténer deux champs pour en obtenir un seul. En définitive, un peu de redondance pour les données de référence peut s'avérer une stratégie payante !

ID	Type de problème	Exemple
112	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 ?

Ici encore, en temps normal, pour toutes les raisons précisées plus haut, il n'est pas utile d'ajouter une colonne dont la valeur est obtenue par calcul à partir d'autres données.

En y regardant de plus près, il existe de nombreux cas dans lesquels cette pratique a un sens : pour de très grosses bases notamment où il est moins coûteux de rajouter un champ calculé plutôt que de

faire un calcul. La performance et la durée du calcul est également un vrai vecteur qui incite à un peu de donnée calculée : il faut que le calcul reste dans un temps acceptable. Et dans beaucoup de cas, le stockage de données pré-calculées est moins coûteux que le calculs et recalculs fréquents que peuvent induire certains usages.

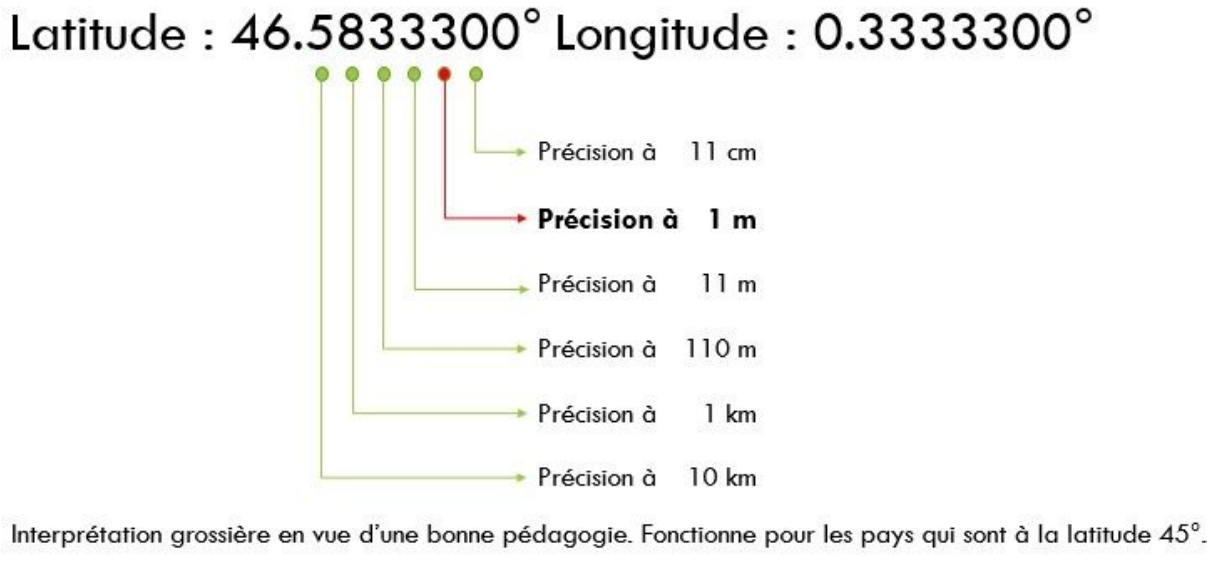
ID	Type de problème	Exemple
113	Surabondance de données en matière de : précision, fréquence, maillage ou fraicheur	La localisation au mm d'une porte d'entrée ; ajout du fuseau horaire dans contexte où toutes les données sont en France métropolitaine ; etc.

Exemple de montant financiers : parfois on a 4 ou 5 chiffres après la virgule ce qui n'est pas justifié. Dans les bases décisionnel d'un assureur deux chiffres après la virgule suffisent. Dans certains cas, une précision à 3 décimales est justifié comme pour le prix de l'essence.

ID	Type de problème	Exemple
114	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

Nous croyons utile de traiter à part ce cas particulier de la surabondance en matière de précision, car il est probablement un des plus fréquents et qu'il est peu visible "à l'oeil nu" pour la plupart des usagers.

Pour des coordonnées géographiques, une précision supérieure à 8 unités après la virgule est inutile
* Exemple : Latitude de 43.29412729123



Interprétation grossière en vue d'une bonne pédagogie. Fonctionne pour les pays qui sont à la latitude 45°.

Source : *Recommandations pour favoriser l'interopérabilité des données open data*, par OpenDataFrance : http://frama.link/ODF-reco-interop

OpenStreetMap, par exemple, limite ses coordonnées à 7 chiffres après la virgule.
À la MAIF par exemple, le géocodeur est paramétré pour sortir des données au mètre.

Au-delà de 8 chiffres après la virgule, on entre donc dans un champ où la marge d'erreur des appareils de mesure est supérieure à 100% du dernier chiffre après la virgule.

Christian Quest : “Lorsqu'on a un trop grand nombre de décimales, ces chiffres souvent peu répétitifs font que les fichiers se compressent mal. Sur certain, en divisant par 2 le nombre de décimales, le fichier non compressé gagnait un peu en taille, mais sa version compressée gagnait beaucoup plus car l'essentiel de l'entropie du fichier venait de ces décimales superflues le reste étant assez répétitif. Sur certains projets, comme la Base Adresse Nationale, l'effet est non négligeable !”

ID	Type de problème	Exemple
119	La durée de conservation des données dépasse une certaine date de péremption ou d’obsolescence (date légale ou date d’usage opérationnel)	Un prospect de plus de 5 ans a des chances de ne plus avoir aucun intérêt opérationnel

Et l’historisation ?
Il y a des politiques liées à des règlementations. Il y a aussi des modes d’historisation lié à l’activité commerciale par exemple.
Exemple des archivistes ?

Point de contrôle sur la date de “péremption” des données ?

- réglementaire
- utile

On garde les informations de parcours client tant qu’on est en relation avec lui mais il faut les supprimer au bout d’un certain temps.

Détection

Sur la question des doublons, il y a une vrai politique vis à vis des doublons.
Chez Carrefour, par exemple, le temps de traitement est particulièrement suivi, qui sert d’alerte sur les problématiques de surabondance de la donnée. Les volumes de stockage sont aussi un indicateur intéressant pour indiquer de possibles surabondance de données.

Corrections

On s’adapte au cas par cas. On revoit le formatage des champs, on réévalue au cas par cas la durée de conservation des données, etc.

6. Les problèmes liés à la réglementation, les conventions d'usage ou à l'éthique [30%]

Ces problèmes sont bien connus des entreprises pour plusieurs raisons. Tout d'abord ils sont anciens et sont donc parfaitement intégrés par les entreprises : les problèmes qualité liés à la loi CNIL, par exemple, sont pris en compte depuis 1978 par le législateur (ID87 à 93). Dotées de données toujours plus nombreuses, les entreprises y font face plus fréquemment.

À ce stade du travail collectif, nous n'avons pas encore eu le temps de les détailler, peut-être aussi parce qu'ils sont justement bien connus. Ils ont néanmoins passé l'épreuve du terrain au cours de nos "sprints qualité". Nous les délivrons ici en l'état.

87	Identification explicite de personnes sans déclaration CNIL	Prénom Nom ou numéro de tél.
88	Identification possible de personnes	Date et lieu de naissance
89	Il existe des jugements de valeurs à propos d'individus	"Client chiant", etc.
90	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	Données d'origine ethnique ou relative à la religion des personnes	"Ne répond pas au téléphone le samedi (shabbat)"
92	Données relatives aux opinions politiques, philosophiques ou à l'appartenance syndicale	"Lié au parti pirate"
93	Données relatives à la vie sexuelle ou au moeurs	"Ménage à 3"
94	Données tierces soumises à licence d'usage	Le fichier publié en Open Data utilise le géocodage de l'API de Google
95	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	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	Données sensibles du point de vu de la sécurité des biens et des personnes	Plan d'une base militaire
98	Données sensibles du point de vu de l'éthique	Localisation de minéraux rares ou de zones d'habitat d'espèces protégées
99	divers : exemple le capital social d'une entreprise s'exprime réglementairement arrondi à la valeur inférieure	

Description du problème qualité et de ses impacts

- La présence de données identifiant assez clairement un individu sans déclaration CNIL

- La présence de données permettant une réidentification de personnes a posteriori
- La présence de données tierces soumises à une licence d'usage, sans respect de cette licence
    - par exemple, des données en open data ne signifient pas que leurs auteurs renoncent à tous droits : dans la plupart des cas, les auteurs demandent a minima une mention d'auteur
    - l'usage de données ouvertes sans mention d'auteur, à tort, est une erreur fréquente
- La présence de données liées à l'origine ethnique ou à la religion des personnes
- La collecte de données sensibles peut poser problème, comme celles liées à la sécurité des biens et des personnes ou de l'État ; typiquement il est interdit de cartographier précisément une centrale nucléaire ou une base militaire
- La collecte de certaines données peut poser des problèmes éthiques au moment de leur restitution ; en soit leur collecte ne pose pas forcément problème, mais leur rediffusion doit être contrôlée
    - typiquement il peut s'agir de la localisation précise d'espèces ou de minéraux rares dont la chasse ou l'exploitation est interdite
    - la cartographie précise des violences à la personne n'est pas interdite en soit, mais leur usage doit poser question ; cette donnée peut aller à l'encontre des politiques publiques en stigmatisant des quartiers
- Certaines données ont un usage réglementaire, comme par exemple le capital social d'une entreprise qui s'exprime arrondi à la valeur entière inférieure