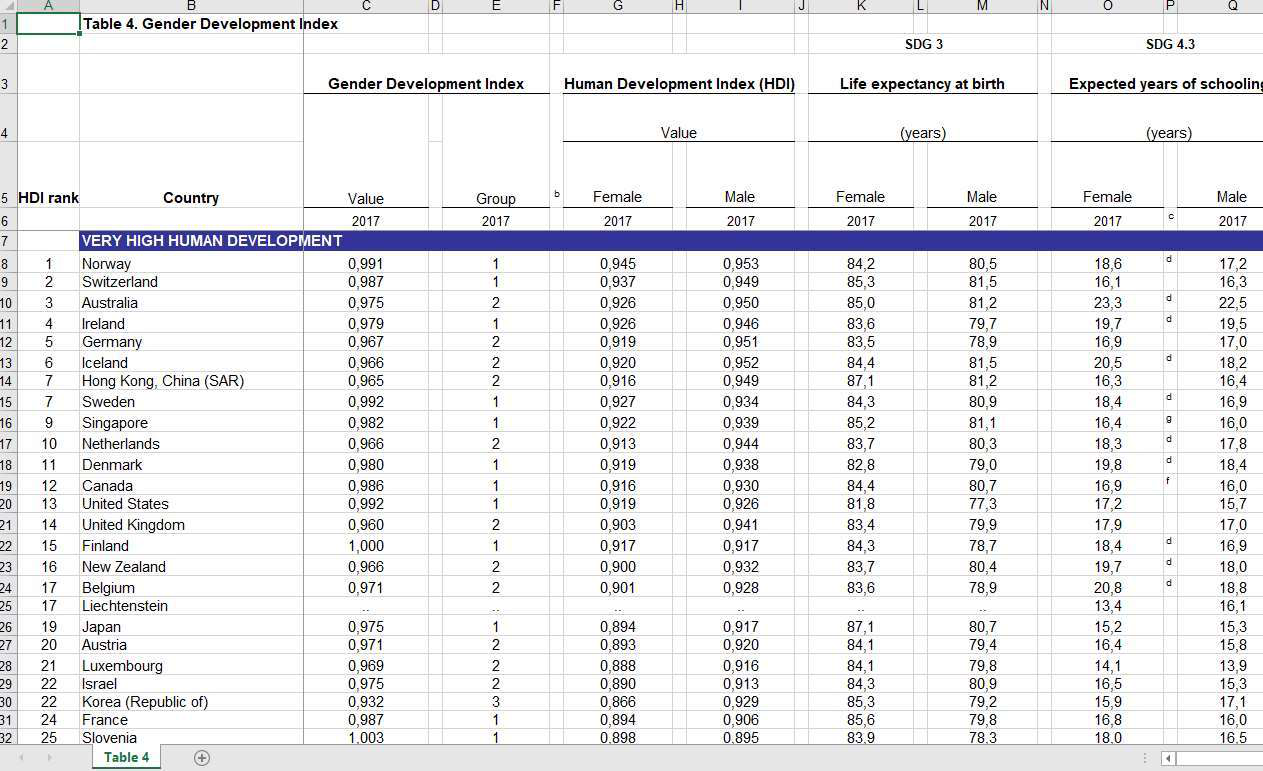




|  |  |  |
| --- | --- | --- |
|  | | |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 7. Bonnes pratiques - 158 | |

|  |
| --- |
|  |
| |  | | --- | | Jeux de données:  Rendre les données lisibles par une machine | |
|  |

|  |  |
| --- | --- |
| Lisible par l’homme |  |



|  |  |  |
| --- | --- | --- |
|  | http://www.sbc4d.com | 7. Bonnes pratiques - 159 |

|  |  |
| --- | --- |
| Lisible par une machine |  |



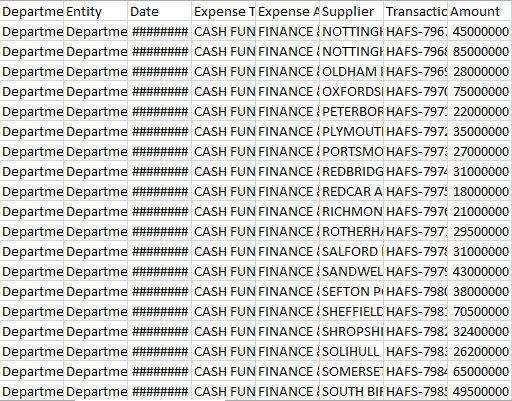
|  |  |  |
| --- | --- | --- |
|  | http://www.sbc4d.com | 7. Bonnes pratiques - 160 |

|  |  |
| --- | --- |
| Données larges |  |



|  |
| --- |
| o Idéal pour la visualisation et le traitement des données o Peu propice au filtrage  o Peu propice à l’archivage  (une mise à jour typique nécessite l’ajout d’une  colonne, donc la modification du schéma de données)  o 16 384 colonnes au maximum dans Excel  http://www.sbc4d.com 7. Bonnes pratiques - 161 |

|  |  |
| --- | --- |
| Données longues |  |



o Filtrage plus facile   
o Idéal pour l’archivage   
o 1 048 576 lignes au maximum dans Excel   
o Le traitement nécessite souvent une transformation préalable… mais elle est facile à faire (et à automatiser)

|  |  |  |
| --- | --- | --- |
|  | http://www.sbc4d.com | 7. Bonnes pratiques - 162 |

|  |  |
| --- | --- |
| Encodage des données |  |

o Problèmes d’encodage classiques:  
 « Où sont les caractères accentués? » « OÃ¹ sont les caractÃ¨res accentuÃ©s? »o Tout encoder en UTF-8!

https://www.w3.org/International/questions/qa-choosing-encodings

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 7. Bonnes pratiques - 163 | |





|  |
| --- |
| Types et valeurs:  Expliciter l’implicite |
|  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 7. Bonnes pratiques - 164 | |

|  |  |
| --- | --- |
| Texte |  |

o Toujours préciser la langue   
Ex: arabe, français, chleuh, anglais

o Normaliser les espaces, ponctuations, accents, majuscules pour favoriser le croisement de données, ex:  
- Écureuil de Barbarie  
- Ecureuil de Barbarie  
- écureuil de Barbarie

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 7. Bonnes pratiques - 165 | |

|  |  |
| --- | --- |
| Nombres |  |
| o Séparateur de décimales: virgule « , » ou point « . »  o Séparateur de milliers:  aucun, espace, ou virgule « , » | Exemples:  5,320.87 5 320,87 5320.87 |

|  |  |  |
| --- | --- | --- |
| o Conventions nationales (suivies par Excel, hélas!)  o vs. Normes internationales  IEEE 754-2019  (JSON, XML, CSV « propre ») | |  |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 7. Bonnes pratiques - 166 | |

|  |  |
| --- | --- |
| Dates |  |

Exemples:

o 12 Mai 2022 à 14:57   
o 12/05/2022 2:57pm   
o 2022-05-12T14:57:00Z

o Utiliser la norme ISO 8601   
o Toujours préciser le fuseau horaire o Privilégier l’heure UTC

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 7. Bonnes pratiques - 167 | |

|  |  |
| --- | --- |
| Numéros de téléphones |  |

o Numéro local, national, international o Norme ITU E.164



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 7. Bonnes pratiques - 168 | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coordonnées GPS | |  | | |
| o  o  o | Minutes d’arc ou décimales?  1° 58' 28,7436" (S) vs -1.974651 Système de coordonnées utilisé?  WGS 84  Interprétation des coordonnées groupées?  « longitude et latitude » ou « latitude et longitude »? | | | |
| Ex: 33.9693414,-6.9273026 | | | | |
|  | | |  | |
| |  | | --- | |  | | | http://www.sbc4d.com | | |  | | --- | | 7. Bonnes pratiques - 169 | |

|  |  |
| --- | --- |
| Données atomiques |  |

o L’interprétation d’une donnée groupée peut prêter à confusion (c.f. coordonnées GPS)   
o Le type d’une donnée groupée est plus difficile à valider (mix entre plusieurs types)   
o Recommandation: Une donnée par colonne

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 7. Bonnes pratiques - 170 | |

|  |  |
| --- | --- |
| Unités et systèmes de référence |  |

o Toujours préciser les unités des valeurs   
numériques (kg, mètre, seconde, etc.)   
o Toujours préciser les systèmes de référence Ex: Calendrier grégorien, calendrier hégirien Ex: WGS84 pour des coordonnées GPS   
Ex: Fuseau horaire, langue   
o Privilégier le système international d’unités (SI) o Privilégier des systèmes de référence communs

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 7. Bonnes pratiques - 171 | |

|  |  |
| --- | --- |
| Valeurs nulles |  |

o Si une valeur dans un jeu de données n’est pas attribuée, laisser le champ vide pour ne pas dénaturer le sens de la donnée.

Ex: colonne numérique, ne pas remplacer par 0 une valeur nulle, sinon les calculs statistiques seront faux

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 7. Bonnes pratiques - 172 | |

|  |  |
| --- | --- |
| Niveau de qualité minimale |  |

o Disponible sous un format ouvert   
 o CSV pour les formats tabulaires (RFC 4180), encodé en UTF8, avec une ligne d’en-têtes, des données numériques au format IEEE 754-

2019, champs séparés par une virgule, lignes délimitées par le

caractère « CRLF »

o TXT, RTF ou HTML pour les fichiers textes   
 o XML ou JSON pour les formats hiérarchiques   
 o PNG, TIFF, ou JPEG pour les images bitmap   
 o SVG pour les images vectorielles   
 o GeoJSON, geotff, shapefile pour les données spatiales o Structurées et exploitables par une machine o Données brutes et désagrégées   
o Données à jour au regard du cycle de collecte o Métadonnées documentées

|  |  |  |
| --- | --- | --- |
|  | http://www.sbc4d.com | 7. Bonnes pratiques - 173 |

|  |  |
| --- | --- |
| Niveau de qualité optimale |  |

o Contraintes du niveau de qualité minimale respectées o Format spécifique au domaine, par exemple:   
 o OCDS pour les marchés publics   
 o GTFS pour les données

o Format long utilisé pour les données tabulaires   
o Valeurs standardisées:   
 o Dates au format ISO 8601   
 o Numéros de téléphone au format ITU E.164   
 o Les métadonnées précisent la langue des champs textes   
 o Les valeurs nulles de champs doivent être vides   
 o Sauf exception, coordonnées géographiques en degrés décimaux, WGS84 (GPS), latitude et longitude séparées en 2 colonnes

o Référentiels utilisés dès que possible

|  |  |  |
| --- | --- | --- |
|  | http://www.sbc4d.com | 7. Bonnes pratiques - 174 |

|  |  |
| --- | --- |
| Niveau de qualité optimale |  |

o Schéma de validation disponible pour les données tabulaires   
o Informations sur la collecte de données   
 disponibles   
o Données « vérifiées »  
 o Validation formelle vis-à-vis du schéma de validation

o Analyses courantes et tests documentés

o Vérification de la complétude des données: toutes les données

historiques sont fournies. Chaque série est complète.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 7. Bonnes pratiques - 175 | |



8. Anonymisation des données

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | 176 |

|  |  |
| --- | --- |
| Données à caractère personnel |  |

« toute information, de quelque nature qu'elle soit et   
indépendamment de son support, y compris le son et l'image, concernant une personne physique identifiée ou   
identifiable, dénommée ci-après personne concernée.

Est réputée identifiable une personne qui peut être   
identifiée, directement ou indirectement, notamment par référence à un numéro d'identification[…] »

Loi n°09-08 du 18 février 2009, relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 8. Anonymisation - 177 | |

|  |  |
| --- | --- |
| Champ d’application de la loi |  |

« La présente loi s'applique au traitement des données à caractère personnel, automatisé en tout ou en partie, ainsi qu’au traitement non automatisé de données à caractère personnel contenues ou appelées à figurer dans des fichiers manuels »

Loi n°09-08 du 18 février 2009, relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 8. Anonymisation - 178 | |

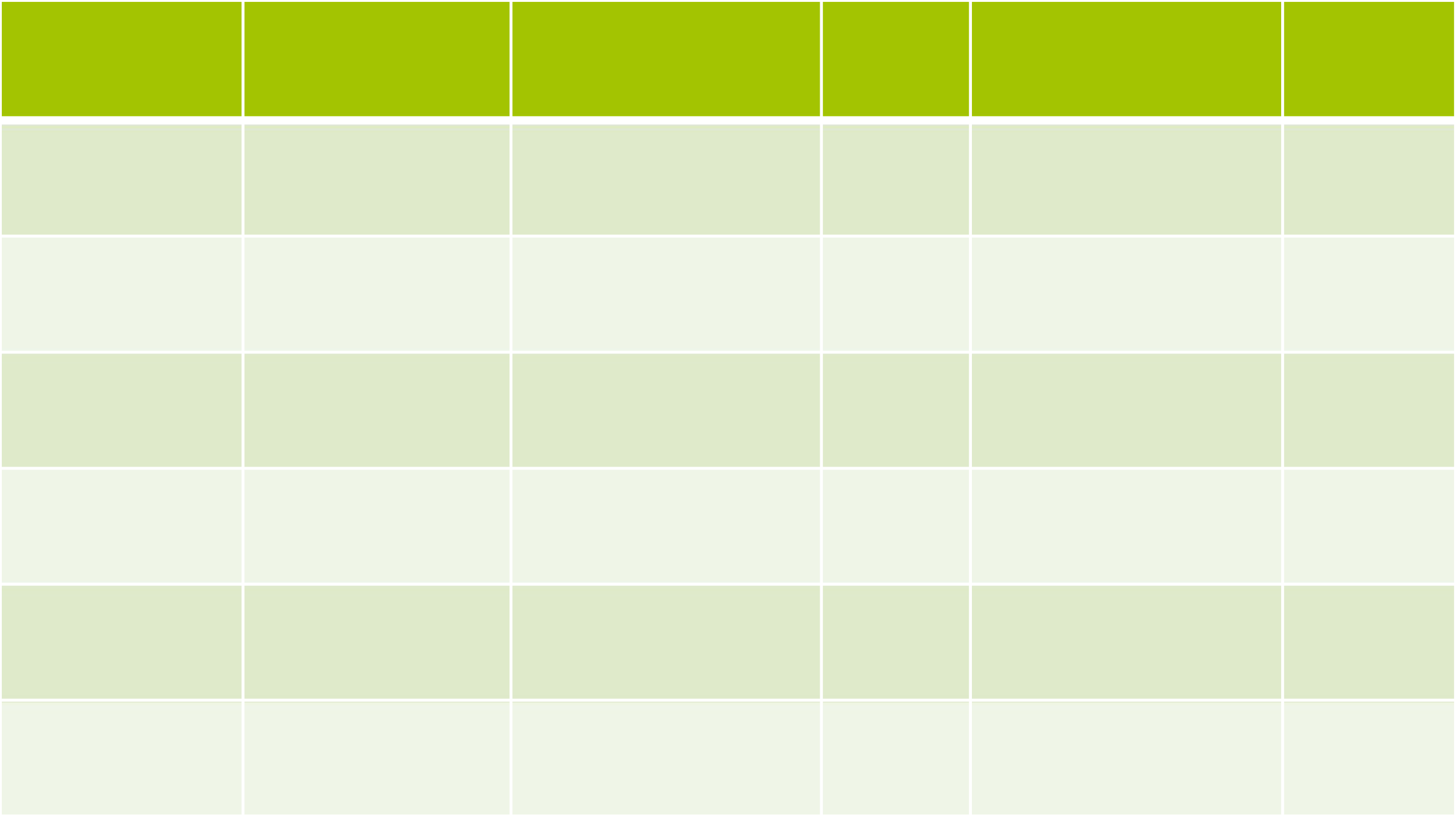
|  |  |
| --- | --- |
| Guides sur l’anonymisation |  |

1. UK ICO – Anonymisation:   
managing data protection risk   
code of practice   
https://ico.org.uk/media/1061/anonymisation-code.pdf

2. EU Commission:   
Opinion 05/2014 on Anonymisation Techniques   
https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216\_en.pdf

3. Singapore:   
Personal Data Protection Commission   
Guide to basic data anonymisation techniques   
https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation\_v1-(250118).pdf

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 8. Anonymisation - 179 | |



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Techniques d’anonymisation | | | | |  |  |
| Nom | Age | Code postal | Sexe | Professeur | Score |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Pierre | 25 | 94110 | M | François | 97 |  |
| Paul | 62 | 75010 | M | Stéphane | 17 |
| Jacques | 61 | 02820 | M | Stéphane | 89 |
| Caroline | 37 | 84250 | F | François | 42 |
| Lisa | 23 | 06410 | F | François | 74 |
| Louise | 40 | 69100 | F | Stéphane | 50 |
| |  | | --- | | … | | … | … | … | … | … |  |
|  |  |  | http://www.sbc4d.com | |  | 8. Anonymisation - 180 |

|  |  |
| --- | --- |
| Techniques d’anonymisation |  |

o Supprimer un attribut   
o Supprimer un enregistrement   
o Masquer des caractères   
o Remplacer les identifiants par des pseudonymes o Généraliser les données   
o Mélanger les données   
o Perturber les données   
o Générer des données synthétiques   
o Agréger les données

(c.f. guide de Singapour pour des exemples)

|  |  |  |
| --- | --- | --- |
|  | http://www.sbc4d.com | 8. Anonymisation - 181 |

|  |  |
| --- | --- |
| Méthodologie |  |

1. Déterminer le type de publication   
2. Déterminer le niveau de risque de ré-identification   
acceptable   
3. Classifier les attributs dans le jeu de données   
4. Supprimer les attributs inutiles   
5. Anonymiser les attributs directs et indirects   
6. Déterminer le risque et comparer avec le risque acceptable 7. Améliorer l’anonymisation, si nécessaire   
8. Evaluer la solution   
9. Déterminer les moyens de contrôles éventuels   
10.Documenter le processus d’anonymisation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 8. Anonymisation - 182 | |

|  |  |
| --- | --- |
| Notion de « k-anonymat » |  |

o Un jeu de données est « k-anonyme » si les identifiants directs et indirects d’un enregistrement sont partagés par au

moins k-1 enregistrements

oLa probabilité de réidentification d’un enregistrement dans un jeu de données k-anonymisé est inférieur à 1/k

oRisques et probabilités maximales typiques: o Risque faible – 0.2 –k >= 5   
o Risque moyen – 0.1 –k >= 10   
o Risque fort – 0.01 –k >= 100

Notions plus avancées: l-diversité / t-proximité

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 8. Anonymisation - 183 | |

|  |  |
| --- | --- |
| Risque de réidentification & données publiques |  |

oTout jeu de données peut être croisé avec d’autres jeux de

données, ce qui accroît le risque de réidentification

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 8. Anonymisation - 184 | |

|  |  |
| --- | --- |
| Importance de l’anonymisation |  |

o Les traitements d’anonymisation doivent avoir lieu AVANT la publication.

o Toutes les données ne nécessitent pas une anonymisation.

o Dans de nombreux cas, les techniques simples de rédaction suffisent.

o Le traitement des cas où les techniques d’agrégation s’imposent requiert une

expertise spécifique qu’il n’est pas

forcément nécessaire de développer au

sein de chaque structure.



|  |  |  |
| --- | --- | --- |
|  | http://www.sbc4d.com | 8. Anonymisation - 185 |

|  |  |
| --- | --- |
| Moyens de contrôle |  |

oRévoquer l’accès aux jeux de données après usage   
o Ne pas publier les données, ne permettre que des requêtes sur les données

oLimiter l’accès à un public restreint   
o Contrôles DRM (empêcher sauvegarde, impression, etc.) oForcer l’accès in-situ   
o Publier uniquement un sous-ensemble du jeu de données

=> Ne s’appliquent pas à une initiative open data!

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 8. Anonymisation - 186 | |

|  |  |
| --- | --- |
| Anonymisation et données géospatiales |  |

Pas de recette miracle pour les données géospatiales, il y a toujours un équilibre à trouver entre:

1. Publier des données aussi désagrégées que possible (à l’échelle d’un quartier, d’une ville, d’un département, par bloc de 100m de côté, etc.)

2. Préserver l’anonymat, en faisant en sorte que chaque cluster contienne un certain nombre de personnes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 8. Anonymisation - 187 | |



9. Validation des données

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | 188 |

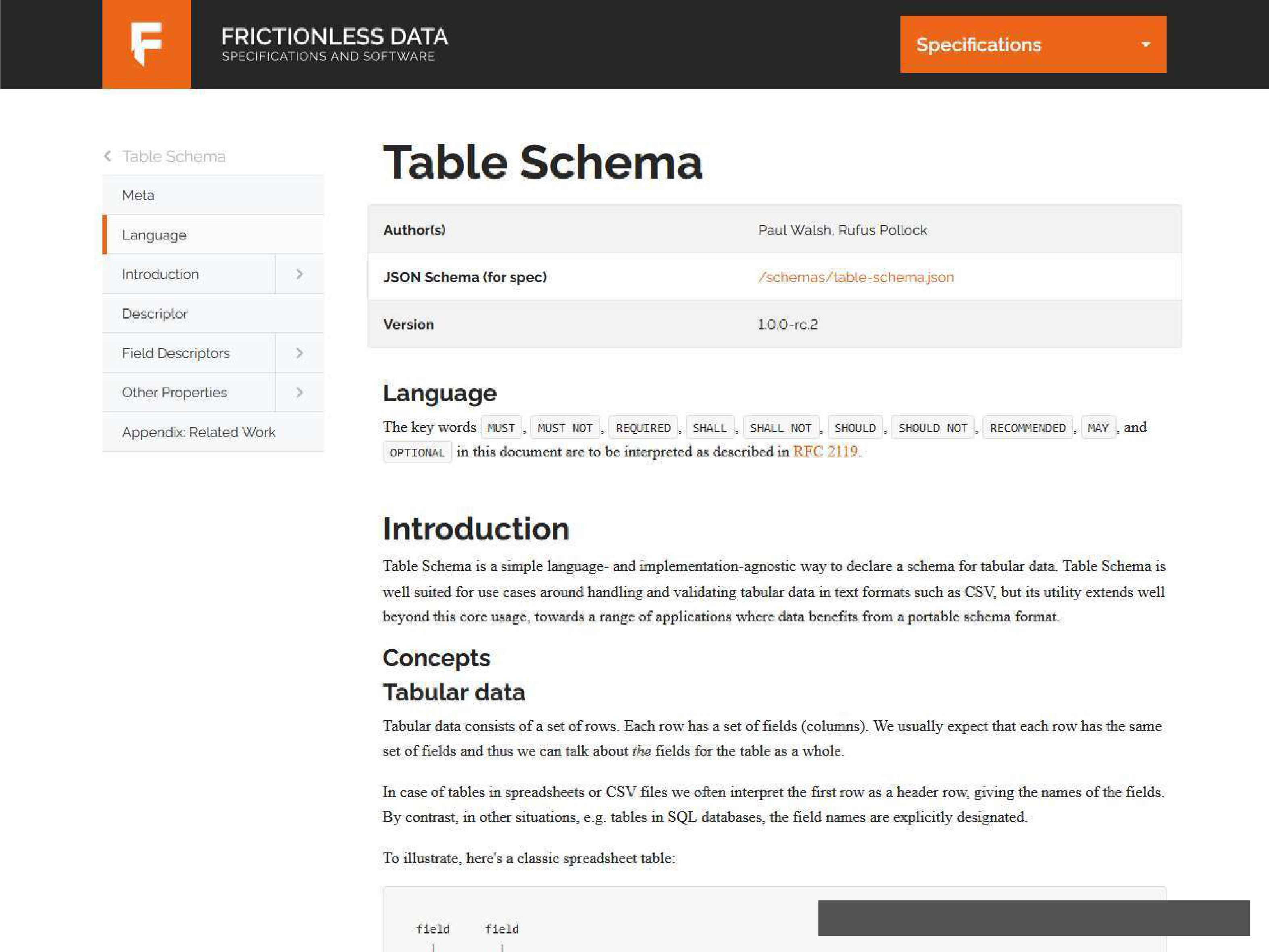
|  |  |
| --- | --- |
| Validation de données |  |

Processus:   
1. Déterminer le format du fichier de données à valider: CSV / JSON / XML / …  
2. Déterminer le format de schéma à utiliser:   
 CSVS / Table Schema / JSON Schema / XML Schema (le schéma dépend du format du fichier de données!) 3. Rédiger le schéma pour le fichier de données   
4. Trouver un validateur adapté

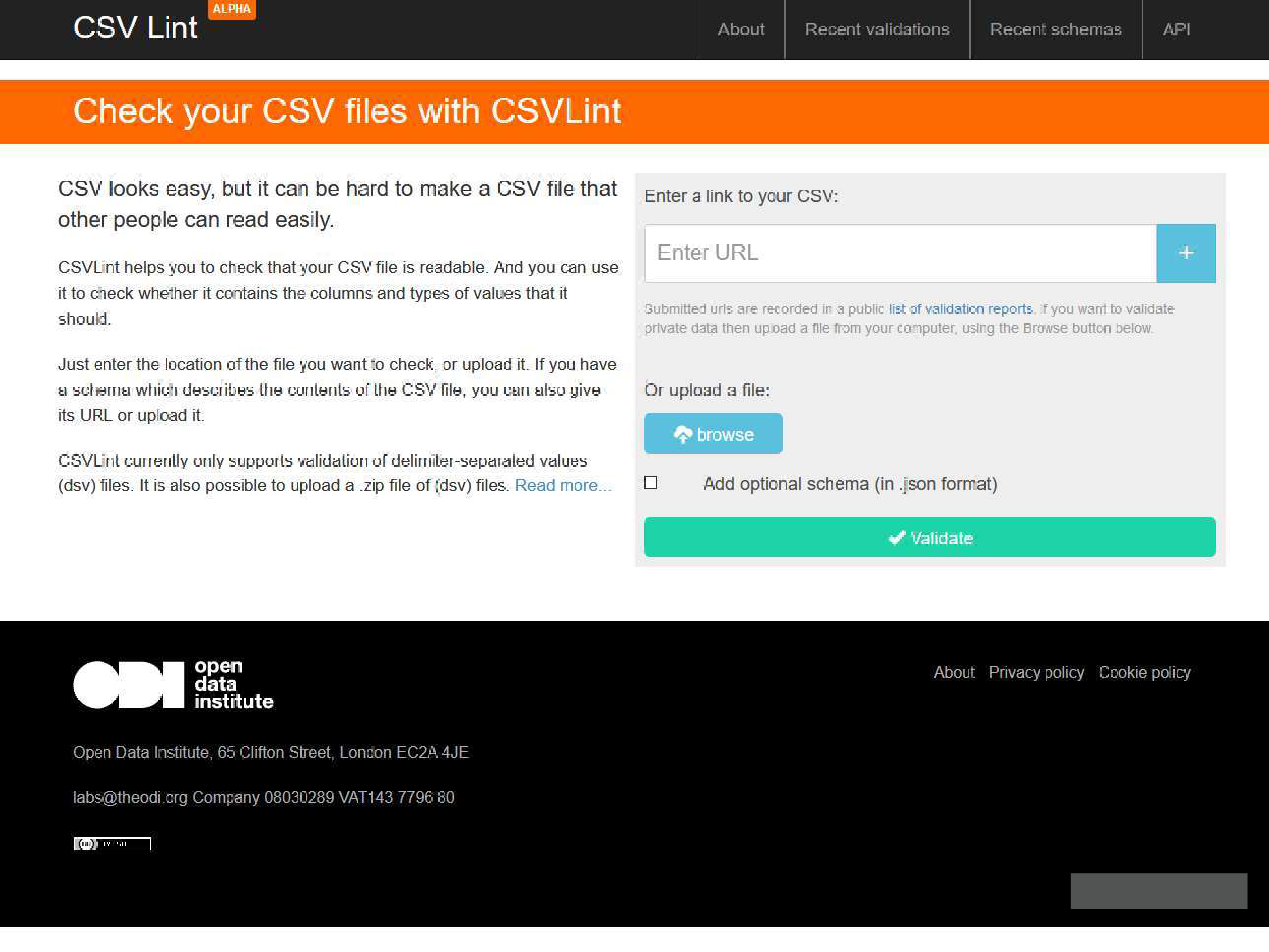
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | |  | | --- | | 9. Validation - 189 | |

|  |  |
| --- | --- |
| Validation de données |  |

|  |  |  |
| --- | --- | --- |
|  | http://www.sbc4d.com | 9. Validation - 190 |



|  |  |
| --- | --- |
| |  | | --- | | https://frictionlessdata.io/specs/table-schema/ | |



|  |  |
| --- | --- |
| |  | | --- | | http://csvlint.io/ | |



Jour 1 – FIN

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | |  | | http://www.sbc4d.com | 193 |