Jeux de données:
Rendre les données lisibles par une machine

Lisible par l'homme

		Table 4. Gender Development Index													
										SDG 3				SDG 4.3	
			Gender Development Index			Human Development Index (HDI)				Life expectancy at birth				Expected years of schooling	
						Value				(years)				(years)	
HDI rank		Country	Value		Group	[b]	Female		Male	Female		Male		Female		Male
			2017		2017		2017		2017	2017		2017		2017	[c]	2017
		VERY HIGH HUMAN DEVELOPMENT													
1		Norway	0,991		1		0,945		0,953	84,2		80,5		18,6	[d]	17,2
2		Switzerland	0,987		1		0,937		0,949	85,3		81,5		16,1		16,3
3		Australia	0,975		2		0,926		0,950	85,0		81,2		23,3	[d]	22,5
4		Ireland	0,979		1		0,926		0,946	83,6		79,7		19,7	[d]	19,5
5		Germany	0,967		2		0,919		0,951	83,5		78,9		16,9		17,0
6		Iceland	0,966		2		0,920		0,952	84,4		81,5		20,5	[d]	18,2
7		Hong Kong, China (SAR)	0,965		2		0,916		0,949	87,1		81,2		16,3		16,4
7		Sweden	0,992		1		0,927		0,934	84,3		80,9		18,4	[d]	16,9
9		Singapore	0,982		1		0,922		0,939	85,2		81,1		16,4	[g]	16,0
10		Netherlands	0,966		2		0,913		0,944	83,7		80,3		18,3	[d]	17,8
11		Denmark	0,980		1		0,919		0,938	82,8		79,0		19,8	[d]	18,4
12		Canada	0,986		1		0,916		0,930	84,4		80,7		16,9	[f]	16,0
13		United States	0,992		1		0,919		0,926	81,8		77,3		17,2		15,7
14		United Kingdom	0,960		2		0,903		0,941	83,4		79,9		17,9		17,0
15		Finland	1,000		1		0,917		0,917	84,3		78,7		18,4	[d]	16,9
16		New Zealand	0,966		2		0,900		0,932	83,7		80,4		19,7	[d]	18,0
17		Belgium	0,971		2		0,901		0,928	83,6		78,9		20,8	[d]	18,8
17		Liechtenstein	..		..		..		..	..		..		13,4		16,1
19		Japan	0,975		1		0,894		0,917	87,1		80,7		15,2		15,3
20		Austria	0,971		2		0,893		0,920	84,1		79,4		16,4		15,8
21		Luxembourg	0,969		2		0,888		0,916	84,1		79,8		14,1		13,9
22		Israel	0,975		2		0,890		0,913	84,3		80,9		16,5		15,3
22		Korea (Republic of)	0,932		3		0,866		0,929	85,3		79,2		15,9		17,1
24		France	0,987		1		0,894		0,906	85,6		79,8		16,8		16,0
25		Slovenia	1,003		1		0,898		0,895	83,9		78,3		18,0		16,5

Table 4

Lisible par une machine

SBC4D

	A	B	C	D	E	F	G	H	I	J	K	L	M
	HDI Rank (2017)	Country	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
	168	Afghanistan											
	68	Albania	0.645	0.626	0.610	0.613	0.619	0.632	0.641	0.641	0.652	0.662	0.669
	85	Algeria	0.577	0.581	0.587	0.591	0.595	0.600	0.608	0.617	0.627	0.636	0.644
	35	Andorra											0.759
	147	Angola										0.374	0.387
	70	Antigua and Barbuda											
	47	Argentina	0.704	0.713	0.720	0.725	0.728	0.731	0.738	0.746	0.753	0.764	0.771
	83	Armenia	0.631	0.628	0.580	0.588	0.599	0.605	0.612	0.623	0.636	0.642	0.647
0	3	Australia	0.866	0.867	0.868	0.872	0.875	0.883	0.886	0.889	0.892	0.895	0.898
1	20	Austria	0.795	0.800	0.805	0.807	0.813	0.817	0.820	0.824	0.835	0.834	0.838
2	80	Azerbaijan						0.612	0.612	0.617	0.626	0.633	0.640
3	54	Bahamas											0.776
4	43	Bahrain	0.746	0.752	0.757	0.765	0.769	0.775	0.778	0.779	0.783	0.786	0.792
5	136	Bangladesh	0.387	0.394	0.402	0.409	0.417	0.425	0.433	0.442	0.451	0.460	0.468
6	58	Barbados	0.716	0.718	0.718	0.721	0.727	0.731	0.735	0.740	0.736	0.743	0.752
7	53	Belarus						0.657	0.661	0.667	0.671	0.676	0.683
8	17	Belgium	0.806	0.810	0.825	0.839	0.845	0.852	0.857	0.862	0.866	0.868	0.873
9	106	Belize	0.644	0.651	0.657	0.661	0.661	0.662	0.662	0.664	0.666	0.670	0.677
0	163	Benin	0.348	0.354	0.358	0.365	0.368	0.373	0.377	0.381	0.385	0.391	0.398
1	134	Bhutan											
2	118	Bolivia (Plurinational State of)	0.536	0.543	0.550	0.557	0.564	0.571	0.578	0.580	0.591	0.600	0.608
3	77	Bosnia and Herzegovina											0.672
4	101	Botswana	0.581	0.588	0.587	0.585	0.575	0.577	0.571	0.570	0.567	0.566	0.565
5	79	Brazil	0.611	0.615	0.622	0.630	0.640	0.648	0.656	0.664	0.670	0.676	0.684
6	39	Brunei Darussalam	0.782	0.787	0.792	0.797	0.801	0.805	0.807	0.810	0.812	0.818	0.819
7	51	Bulgaria	0.694	0.691	0.691	0.690	0.691	0.696	0.702	0.704	0.709	0.708	0.712
8	183	Burkina Faso											0.286
9	185	Burundi	0.297	0.300	0.297	0.299	0.298	0.296	0.294	0.296	0.300	0.300	0.303
0	125	Cabo Verde											0.570
1	146	Cambodia	0.364	0.368	0.373	0.377	0.380	0.387	0.391	0.397	0.402	0.407	0.420
2	151	Cameroon	0.440	0.436	0.432	0.426	0.423	0.422	0.422	0.422	0.428	0.426	0.431
3	12	Canada	0.849	0.853	0.856	0.854	0.859	0.861	0.863	0.862	0.861	0.864	0.867
4	188	Central African Republic	0.317	0.312	0.299	0.299	0.301	0.303	0.300	0.302	0.304	0.308	0.309
5	186	Chad											0.299
6	44	Chile	0.701	0.711	0.719	0.713	0.718	0.727	0.734	0.741	0.747	0.753	0.759

Données larges

Country Name	Country Code	Indicator Name	Indicator Code	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
Aruba	ABW	Urban populatio	SP.URB.TOTL	27526	28141	28532	28761	28924	29082	29253	29416	29575	29738
Afghanistan	AFG	Urban populatio	SP.URB.TOTL	755836	796272	839385	885228	934135	986074	1041191	1099272	1161355	1228273
Angola	AGO	Urban populatio	SP.URB.TOTL	569222	597288	628381	660180	691532	721552	749534	776116	804107	837758
Albania	ALB	Urban populatio	SP.URB.TOTL	493982	513592	530766	547928	565248	582374	599300	616687	635924	656733
Andorra	AND	Urban populatio	SP.URB.TOTL	7839	8766	9754	10811	11915	13067	14262	15494	16765	18083
Arab World	ARB	Urban populatio	SP.URB.TOTL	28797177	30292822	31856717	33513046	35275337	37163923	39098493	41001112	42996408	45072707
United Arab Emirates	ARE	Urban populatio	SP.URB.TOTL	67927	74975	84367	95215	106178	116473	125594	134581	145736	162079
Argentina	ARG	Urban populatio	SP.URB.TOTL	15076842	15449950	15815502	16183085	16552517	16923103	17295214	17669088	18048312	18436396
Armenia	ARM	Urban populatio	SP.URB.TOTL	960956	1012430	1065431	1119586	1174560	1229980	1285572	1341279	1397345	1454162
American Samoa	ASM	Urban populatio	SP.URB.TOTL	13324	13729	14254	14871	15522	16176	16818	17462	18082	18687
Antigua and Barbuda	ATG	Urban populatio	SP.URB.TOTL	21466	21472	21458	21443	21449	21489	21577	21690	21775	21763
Australia	AUS	Urban populatio	SP.URB.TOTL	8378309	8589875	8832932	9034955	9245383	9459784	9709943	9852991	10048170	10280809
Austria	AUT	Urban populatio	SP.URB.TOTL	4561167	4592914	4624644	4658034	4692726	4726878	4763809	4803163	4831802	4852163
Azerbaijan	AZE	Urban populatio	SP.URB.TOTL	2051433	2110438	2171811	2233682	2293589	2349762	2401555	2449253	2493161	2534087
Burundi	BDI	Urban populatio	SP.URB.TOTL	58113	60329	62624	65010	67577	70985	75933	81363	87127	93063
Belgium	BEL	Urban populatio	SP.URB.TOTL	8463316	8500111	8545539	8624158	8720520	8814176	8887919	8951328	9000174	9039007
Benin	BEN	Urban populatio	SP.URB.TOTL	225533	243036	262053	282715	305170	329545	356057	384830	416031	449720
Burkina Faso	BFA	Urban populatio	SP.URB.TOTL	226977	234744	242709	251039	259788	268990	278745	289111	299959	311347
Bangladesh	BGD	Urban populatio	SP.URB.TOTL	2465493	2605371	2790354	2989609	3205156	3439969	3696385	3974776	4269774	4573116

- Idéal pour la visualisation et le traitement des données
- Peu propice au filtrage
- **Peu propice à l'archivage**
  (une mise à jour typique nécessite l'ajout d'une colonne, donc la modification du schéma de données)
- 16 384 colonnes au maximum dans Excel

Données longues

Departme	Entity	Date	Expense T	Expense A	Supplier	Transactic	Amount
Departme	Departme	########	CASH FUN	FINANCE &	NOTTINGH	HAFS-7961	45000000
Departme	Departme	########	CASH FUN	FINANCE &	NOTTINGH	HAFS-7968	85000000
Departme	Departme	########	CASH FUN	FINANCE &	OLDHAM I	HAFS-7969	28000000
Departme	Departme	########	CASH FUN	FINANCE &	OXFORDSI	HAFS-7970	75000000
Departme	Departme	########	CASH FUN	FINANCE &	PETERBOR	HAFS-7971	22000000
Departme	Departme	########	CASH FUN	FINANCE &	PLYMOUTI	HAFS-7972	35000000
Departme	Departme	########	CASH FUN	FINANCE &	PORTSMO	HAFS-7973	27000000
Departme	Departme	########	CASH FUN	FINANCE &	REDBRIDG	HAFS-7974	31000000
Departme	Departme	########	CASH FUN	FINANCE &	REDCAR A	HAFS-7975	18000000
Departme	Departme	########	CASH FUN	FINANCE &	RICHMON	HAFS-7976	21000000
Departme	Departme	########	CASH FUN	FINANCE &	ROTHERH/	HAFS-7977	29500000
Departme	Departme	########	CASH FUN	FINANCE &	SALFORD	HAFS-7978	31000000
Departme	Departme	########	CASH FUN	FINANCE &	SANDWEL	HAFS-7979	43000000
Departme	Departme	########	CASH FUN	FINANCE &	SEFTON P	HAFS-7980	38000000
Departme	Departme	########	CASH FUN	FINANCE &	SHEFFIELD	HAFS-7981	70500000
Departme	Departme	########	CASH FUN	FINANCE &	SHROPSHI	HAFS-7982	32400000
Departme	Departme	########	CASH FUN	FINANCE &	SOLIHULL	HAFS-7983	26200000
Departme	Departme	########	CASH FUN	FINANCE &	SOMERSET	HAFS-7984	65000000
Departme	Departme	########	CASH FUN	FINANCE &	SOUTH BIF	HAFS-7985	49500000

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

Encodage des données

SBC4D

- **Problèmes d'encodage classiques:**
  « Où sont les caractères accentués? »
  « OÃ¹ sont les caractÃ¨res accentuÃ©s? »

- **Tout encoder en UTF-8!**

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

Types et valeurs: Expliciter l'implicite

Texte

- **Toujours préciser la langue**
  Ex: arabe, français, chleuh, anglais

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

Nombres

- Séparateur de décimales:
  virgule « , » ou **point** « . »
- Séparateur de milliers:
  aucun, espace, ou virgule « , »

Exemples:
5,320.87
5 320,87
5320.87

- Conventions nationales
  (suivies par Excel, hélas!)
- vs. Normes internationales

IEEE 754-2019
(JSON, XML, CSV « propre »)

Dates

Exemples:

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

- Utiliser la norme **ISO 8601**
- Toujours préciser le **fuseau horaire**
- Privilégier **l'heure UTC**

Numéros de téléphones

- Numéro local, national, international
- **Norme ITU E.164**

Coordonnées GPS

- 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

Données atomiques

SBC4D

- L'interprétation d'une donnée groupée peut prêter à confusion (c.f. coordonnées GPS)
- Le type d'une donnée groupée est plus difficile à valider (mix entre plusieurs types)

- Recommandation: **Une donnée par colonne**

Unités et systèmes de référence

SBC4D

- Toujours préciser les **unités** des valeurs numériques (kg, mètre, seconde, etc.)

- Toujours préciser les **systèmes de référence**
  Ex: Calendrier grégorien, calendrier hégirien
  Ex: WGS84 pour des coordonnées GPS
  Ex: Fuseau horaire, langue

- Privilégier le système international d'unités (SI)
- Privilégier des systèmes de référence communs

Valeurs nulles

- 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

Niveau de qualité minimale

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

Niveau de qualité optimale

- Contraintes du niveau de qualité minimale respectées
- Format spécifique au domaine, par exemple:
  - OCDS pour les marchés publics
  - GTFS pour les données
- Format long utilisé pour les données tabulaires
- Valeurs standardisées:
  - Dates au format ISO 8601
  - Numéros de téléphone au format ITU E.164
  - Les métadonnées précisent la langue des champs textes
  - Les valeurs nulles de champs doivent être vides
  - Sauf exception, coordonnées géographiques en degrés décimaux, WGS84 (GPS), latitude et longitude séparées en 2 colonnes
  - Référentiels utilisés dès que possible

Niveau de qualité optimale

- Schéma de validation disponible pour les données tabulaires
- Informations sur la collecte de données disponibles
- Données « vérifiées »
  - Validation formelle vis-à-vis du schéma de validation
  - Analyses courantes et tests documentés
  - Vérification de la complétude des données: toutes les données historiques sont fournies. Chaque série est complète.

8. Anonymisation des données

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.

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.

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

Techniques d'anonymisation

SBC4D

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
…	…	…	…	…	…

Techniques d'anonymisation

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

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

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

Notion de « k-anonymat »

- Un jeu de données est « **k-anonyme** » si les identifiants directs et indirects d'un enregistrement sont partagés par **au moins k-1** enregistrements
- La probabilité de réidentification d'un enregistrement dans un jeu de données k-anonymisé est inférieur à **1/k**
- Risques et probabilités maximales typiques:
  - Risque faible – 0.2 – **k >= 5**
  - Risque moyen – 0.1 – **k >= 10**
  - Risque fort – 0.01 – **k >= 100**

Notions plus avancées: l-diversité / t-proximité

Risque de réidentification & données publiques

- Tout jeu de données peut être croisé avec d'autres jeux de données, ce qui accroît le risque de réidentification

Importance de l'anonymisation

- Les traitements d'anonymisation doivent avoir lieu **AVANT** la publication.
- Toutes les données ne nécessitent pas une anonymisation.
- Dans de nombreux cas, les techniques simples de rédaction suffisent.
- 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.

Moyens de contrôle

- Révoquer l'accès aux jeux de données après usage
- Ne pas publier les données, ne permettre que des requêtes sur les données
- Limiter l'accès à un public restreint
- Contrôles DRM (empêcher sauvegarde, impression, etc.)
- Forcer l'accès in-situ
- Publier uniquement un sous-ensemble du jeu de données

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

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**

9. Validation des données

Validation de données

SBC4D

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

Validation de données

SBC4D

Données	Schéma	Outil
CSV	Table Schema	https://csvlint.io
		https://github.com/frictionlessdata/tableschema-py
	CSVS	http://digital-preservation.github.io/csv-validator/
JSON	JSON Schema	https://www.jsonschemavalidator.net/
XML	XML Schema	https://www.liquid-technologies.com/online-xsd-validator
GTFS	N/A	https://github.com/google/transitfeed/wiki/FeedValidator

Table Schema

Author(s)	Paul Walsh, Rufus Pollock
JSON Schema (for spec)	/schemas/table-schema.json
Version	1.0.0-rc.2

Language

The key words `MUST`, `MUST NOT`, `REQUIRED`, `SHALL`, `SHALL NOT`, `SHOULD`, `SHOULD NOT`, `RECOMMENDED`, `MAY`, and `OPTIONAL` in this document are to be interpreted as described in RFC 2119.

Introduction

Table Schema is a simple language- and implementation-agnostic way to declare a schema for tabular data. Table Schema is well suited for use cases around handling and validating tabular data in text formats such as CSV, but its utility extends well beyond this core usage, towards a range of applications where data benefits from a portable schema format.

Concepts

Tabular data

Tabular data consists of a set of rows. Each row has a set of fields (columns). We usually expect that each row has the same set of fields and thus we can talk about *the* fields for the table as a whole.

In case of tables in spreadsheets or CSV files we often interpret the first row as a header row, giving the names of the fields. By contrast, in other situations, e.g. tables in SQL databases, the field names are explicitly designated.

To illustrate, here's a classic spreadsheet table:

```
field     field
```

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

Check your CSV files with CSVLint

CSV looks easy, but it can be hard to make a CSV file that other people can read easily.

CSVLint helps you to check that your CSV file is readable. And you can use it to check whether it contains the columns and types of values that it should.

Just enter the location of the file you want to check, or upload it. If you have a schema which describes the contents of the CSV file, you can also give its URL or upload it.

CSVLint currently only supports validation of delimiter-separated values (dsv) files. It is also possible to upload a .zip file of (dsv) files. Read more...

Enter a link to your CSV:

Enter URL +

Submitted urls are recorded in a public list of validation reports. If you want to validate private data then upload a file from your computer, using the Browse button below.

Or upload a file:

browse

☐ Add optional schema (in .json format)

✔ Validate

Open Data Institute, 65 Clifton Street, London EC2A 4JE

labs@theodi.org Company 08030289 VAT143 7796 80

http://csvlint.io/

Jour 1 - FIN