



# Learn Git and GitHub without any code!

Using the Hello World guide, you'll start a branch, write comments, and open a pull request.

[Read the guide](#)

[whythawk](#) / [data-wrangling-and-validation](#)

Branch: master ▼

[data-wrangling-and-validation](#) / [Leçon 1-1 - Structurer et organiser des données désordonnées à l'aide d'un tableur.ipynb](#)

[Find file](#)

[Copy path](#)



**turukawa** Updated tutorial data links

5f3c13b 20 hours ago

[1 contributor](#)

597 lines (597 sloc) 57 KB



[Raw](#)

[Blame](#)

[History](#)



## 1. Structurer et organiser des données désordonnées à l'aide d'un tableur

**A la fin de la formation, vous pourrez:**

- Comprendre et avoir une expérience pratique de la structure et de la conception des fichiers de données exploitables par une machine.
- Utiliser Excel pour étudier et manipuler des données sources afin de comprendre leurs métadonnées, leur forme et leur robustesse, et utiliser ces méthodes pour développer un schéma de métadonnées structurées.
- Apprendre et appliquer un ensemble de méthodes de base pour restructurer

des données source désordonnées en fichiers CSV exploitable par une machine en utilisant Microsoft Excel.

## 1.1 Introduction

L'expression "données ouvertes" s'entend généralement des données qui sont mises à la disposition du public gratuitement, sans enregistrement ni licence restrictive, à quelque fin que ce soit (y compris à des fins commerciales), dans des formats électroniques exploitables par une machine qui garantissent que les données sont faciles à trouver, à télécharger et à utiliser.

Les initiatives en matière de données ouvertes prises par les institutions publiques, telles que les gouvernements et les organisations intergouvernementales, reconnaissent que ces données sont produites avec des fonds publics et doivent donc, à quelques exceptions près, être traitées comme des biens publics.

La réutilisation des données, tant par les experts en données que par le grand public, est essentielle pour créer de nouvelles opportunités et de nouvelles connaissances à partir des données gouvernementales. La réutilisation des données ouvertes requiert deux critères de base :

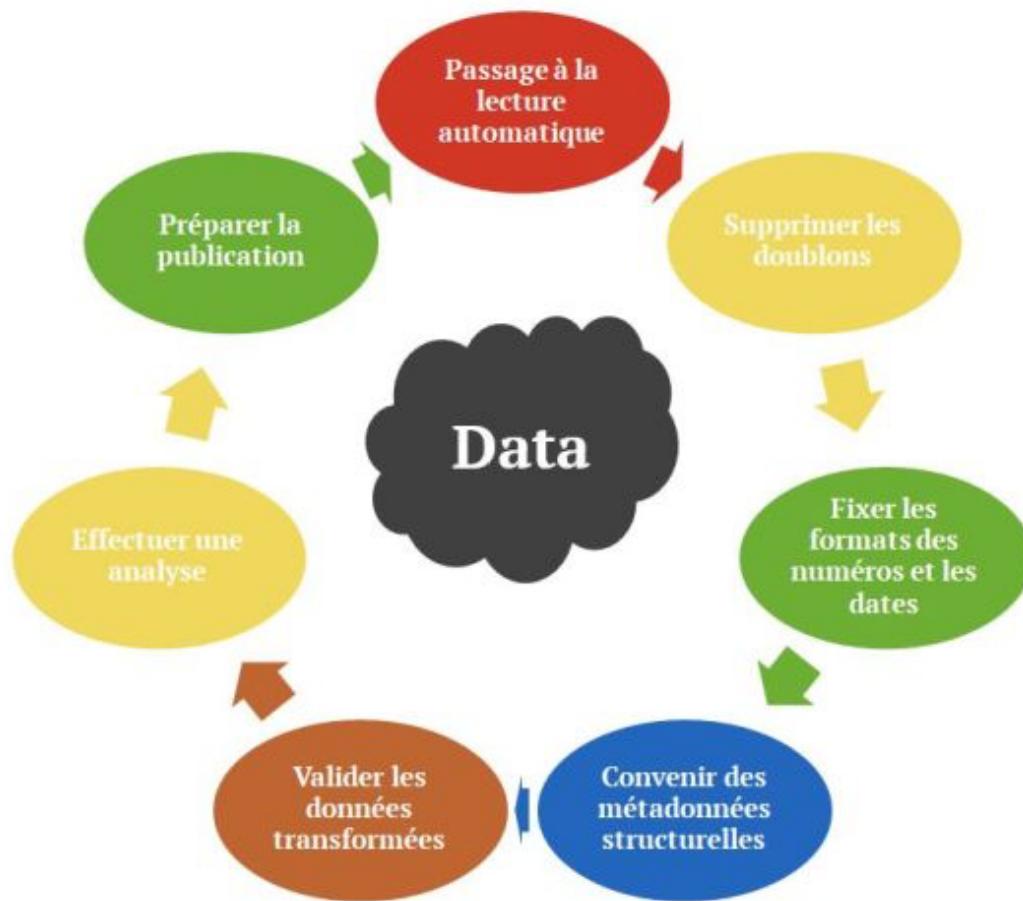
1. **Les données doivent être légalement ouvertes**, ce qui signifie qu'elles sont placées dans le domaine public ou dans des conditions d'utilisation ouverte avec un minimum de restrictions. Cela permet de garantir que les politiques gouvernementales ne créent pas d'obstacles ou d'ambiguïtés quant à la manière dont les données peuvent être utilisées.
2. **Les données doivent être techniquement ouvertes**, ce qui signifie qu'elles sont publiées dans des formats électroniques exploitables par une machine et non propriétaires. Cela garantit que les citoyens ordinaires peuvent accéder aux données et les utiliser à peu de frais, voire gratuitement, en utilisant des outils logiciels courants.

L'objectif de ce programme de formation en *Analyse et validation de données ouvertes* est de guider les participants et les aider à fournir des données techniquement ouvertes : des données bien structurées, exploitables par une machine, validées selon un schéma de métadonnées défini et standard.

## 1.2 Le cycle de vie de la gestion des données

Si le processus de création, de maintenance et d'exploitation de nouvelles données est souvent présenté comme un cycle, il s'agit plutôt d'une spirale. Chaque cycle monte en spirale, et donne lieu à plus d'informations. Cependant, l'efficacité de chaque étape est définie par les

besoins de ses utilisateurs, et sa pertinence par rapport au processus ou aux événements qu'elle reflète.



### 1.2.3 Collecte et création

Avant de pouvoir collecter des données, il faut connaître toute une série de choses :

- pourquoi collectons-nous les données ?
- à quoi servent-elles ?
- avons-nous le consentement et/ou l'autorisation légale de collecter ces données ?
- existe-t-il une série de données existantes que ces données complètent ou élargissent ?
- comment les données seront-elles collectées ?
- qui sera responsable de la qualité des données et comment cette qualité sera-t-elle mesurée ?
- qui aura accès aux données et quel est le degré de sensibilité de ces données (par exemple, identification personnelle) ?
- utilisons-nous une classification ou un format de métadonnées standardisé et convenu ?

### 1.2.4 Classification et traitement

Le créateur des données est le mieux placé pour savoir de quoi il s'agit et devrait attribuer des mots-clés comme descripteurs. Ces données sur les données sont appelées métadonnées. Le

metadonnées des données descriptives. Ces données sur les données sont appelées métadonnées. Le terme est ambigu, car il est utilisé pour deux concepts fondamentalement différents :

- **Les métadonnées structurelles** correspondent aux métadonnées internes (c'est-à-dire les métadonnées concernant la structure des objets de la base de données tels que les tables, les colonnes, les clés et les index) ;
- **Les métadonnées descriptives**, qui correspondent à des métadonnées externes. (c'est-à-dire des métadonnées typiquement utilisées pour la découverte et l'identification, comme des informations utilisées pour rechercher et localiser un objet tel que le titre, l'auteur, les sujets, les mots clés, l'éditeur) ;

Les métadonnées descriptives permettent la découverte de l'objet. Les métadonnées structurelles permettent d'appliquer, d'interpréter, d'analyser, de restructurer les données et de les relier à d'autres ensembles de données similaires.

Les métadonnées peuvent permettre l'interopérabilité entre différents systèmes. Une structure convenue pour interroger la "fiabilité" d'une série de données peut permettre à des systèmes logiciels indépendants de trouver et d'utiliser des données à distance.

Au-delà des métadonnées, il existe également des mécanismes permettant de structurer les relations entre les hiérarchies de mots clés. Ces mécanismes sont connus sous le nom d'ontologies et, avec les métadonnées, peuvent être utilisés pour définir avec précision et permettre la découverte de données.

L'ajout de métadonnées aux ressources de données existantes peut être un processus coûteux et exigeant en main-d'œuvre. Cela peut devenir un obstacle à la mise en œuvre d'un système complet de gestion des connaissances.

Les types de métadonnées descriptives que nous devrions inclure (et les termes que nous devrions utiliser pour les désigner) :

OBLIGATOIRE	RECOMMANDÉ	FACULTATIF
Titre	Tag(s)	Dernière mise à jour
Description	Conditions d'utilisation	Fréquence de mise à jour
Thème(s)	Courriel de contact	Couverture géographique
Organisme de publication		Couverture temporelle
		Validité
		Ressources connexes
		Règlements

## 1.2.5 Manipulation, conversion ou altération

Cette partie du processus est celle où les données sont transcrites, traduites, vérifiées, validées, nettoyées et gérées.

C'est ce qui présente le plus grand risque pour la cohérence des données. Tout changement de format ou toute manipulation, ou même la copie d'un fichier d'un système à un autre, introduit un risque de corruption des données. De même, cela augmente également le risque que des données - qu'elles soient erronées ou non - soient accidentellement communiquées aux utilisateurs ou au public avant qu'elles ne soient prêtes.

C'est ce que l'on appelle le **data wrangling** ou préparation des données (il n'y a pas de traduction parfaite pour ce terme et vous le retrouverez souvent sous forme anglaise).

### 1.2.6 Analyse et présentation

L'analyse ne fait peut-être pas toujours partie du rôle du gestionnaire des données, mais c'est certainement la raison pour laquelle les données sont collectées.

C'est là que les données sont interprétées, combinées à d'autres ensembles de données pour produire une méta-analyse, et que l'analyse devient l'histoire que vous souhaitez raconter à partir des données.

La qualité de l'histoire dépend de celle de vos données et la recherche n'est considérée comme valable que si les données qui l'alimentent sont également publiées.

L'objectif de l'analyse est d'informer le comportement collectif ou individuel, d'influencer la politique ou de soutenir l'activité économique, parmi beaucoup d'autres. La confiance peut être obtenue en "montrant votre fonctionnement" (c'est à dire comment vous avez analysé les données), ce qui inclut la publication des données.

### 1.2.7 Préservation et stockage

Les données doivent être préservées de la corruption et être disponibles pour une utilisation ultérieure une fois l'analyse initiale terminée. Il est également nécessaire de préserver les données en cas de questions concernant l'analyse.

Le stockage à long terme exige que les **métadonnées** soient bien définies et soient vraiment utiles pour garantir que la compréhension de ce que les données décrivent est encore possible longtemps après leur collecte initiale.

Les données peuvent finir par être stockées dans plusieurs formats ou sur plusieurs systèmes. Il est essentiel que la primauté soit établie (c'est-à-dire quel ensemble de données est la version originale et prime sur les autres) et que les différents formats soient maintenus en alignement.

### 1.2.8 Publication et accès

Même lorsque les données ne sont diffusées qu'au sein du gouvernement ou d'une organisation - et non pour le public - il y aura toujours d'autres personnes qui voudront utiliser vos données, ou qui en tireront profit si elles savent qu'elles existent. La plus grande inefficacité dans la gestion des données survient lorsque la recherche est répétée parce qu'un autre ministère, un autre département, une autre équipe de l'organisation a besoin des mêmes

données sans savoir qu'elles existent déjà.

La diffusion ne consiste pas simplement à mettre les données à la disposition du public, mais aussi à créer un processus prévisible pour cette diffusion.

Les données régulièrement collectées (comme les taux d'inflation) ont besoin d'un cycle de diffusion prévisible, car de nombreuses entreprises fondent leurs décisions d'investissement sur la disponibilité de ces informations.

Publier un calendrier de diffusion pour vos consommateurs de données (et s'y tenir) leur permet de planifier leur propre analyse, ou de réagir à votre analyse.

L'accès implique que vous ayez besoin d'une base de données centralisée qui soit accessible à vos utilisateur. Dans le cas de données ouvertes, comment les données seront-elles déplacées des serveurs internes vers un dépôt public ?

La responsabilité de ce processus doit être attribuée et mesurée.

### 1.2.9 Conservation et réutilisation

Une fois les données publiées, la question se pose de savoir pendant combien de temps elles seront disponibles. Les données de recherche devraient, en règle générale, être disponibles à perpétuité.

Les données chronologiques sont d'autant plus utiles qu'elles ont été collectées longtemps. La suppression soudaine de données peut provoquer d'énormes perturbations.

Si des systèmes et un soutien appropriés n'ont pas été mis en place, la conservation peut devenir un problème très coûteux.

Il est important, pour favoriser la réutilisation, de définir clairement les droits d'auteur et les licences qui permettent de réutiliser librement les données à n'importe quelle fin.

### 1.2.10 Archivage

Les ensembles de données peuvent devenir très volumineux et ne peuvent être consultés que rarement. Cela peut devenir problématique pour le stockage à long terme.

Un processus d'archivage - où les données peuvent être stockées à moindre coût tout en restant accessibles - peut devoir être envisagé.

#### Actions:

- S'assurer qu'il y a un propriétaire des données qui est responsable du cycle de vie des données, y compris de leur publication et de la réponse aux commentaires ou aux questions;
- Préparer un plan de collecte de données, en veillant à ce que les définitions et la structure des données soient conformes aux normes internationales;
- S'assurer que les métadonnées sont convenues avec les parties prenantes et

qu'elles sont utiles et normalisées;

- S'assurer que les données font l'objet d'une licence appropriée pour garantir leur diffusion et leur réutilisation;

### Références:

- Cycle de vie des données de l'université de Boston  
(<http://www.bu.edu/datamanagement/background/data-life-cycle/>)
- Tech Target Data Lifecycle Management  
(<http://searchsecurity.techtarget.com/magazineContent/Data-Lifecycle-Management-Model-Shows-Risks-and-Integrated-Data-Flow>)
- Norme de métadonnées de Dublin Core  
([http://en.wikipedia.org/wiki/Dublin\\_Core](http://en.wikipedia.org/wiki/Dublin_Core))
- Norme de métadonnées DCAT (<http://www.w3.org/TR/vocab-dcat/>)
- Sept outils de base de qualité  
([http://en.wikipedia.org/wiki/Seven\\_Basic\\_Tools\\_of\\_Quality](http://en.wikipedia.org/wiki/Seven_Basic_Tools_of_Quality))

## 1.3 Préparation de données et préparation de la publication

Il existe un certain nombre de formats communs pour la distribution des données. Certains sont considérés comme "ouverts" (tels que CSV, XML, texte et autres) et d'autres comme propriétaires (SAS, STATA, SPSS, etc.). XLS et XLSX, associés à Microsoft Excel, sont des formats relativement ouverts et un certain nombre de systèmes logiciels peuvent interpréter les données.

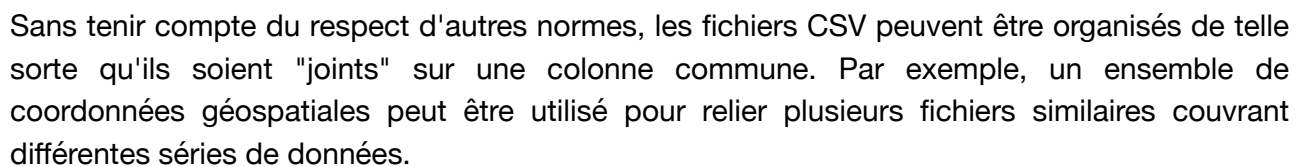
Les formats propriétaires sont légitimes, même dans le cadre d'une initiative de données ouvertes, car ce sont les systèmes logiciels utilisés par de nombreux utilisateurs professionnels de données. Cependant, comme ces formats ne sont souvent pas interopérables, le potentiel de réutilisation des données est limité, à moins que les formats ouverts ne soient également pris en charge. La diffusion de données dans des formats propriétaires n'exclut pas la diffusion dans des formats ouverts et vice versa.

Les tableurs et les systèmes de données distribués ne disposent souvent pas d'une structure de données convenue. Un chercheur qui souhaite combiner ces données avec d'autres données doit d'abord les normaliser, puis décider de termes normalisés pour définir les colonnes et les types de données dans ces colonnes.

AG19															f																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																												
A		B		C		D		E		F		G		H		I		J		K		L		M		N		O		P		Q		R		S		T		U		V		W		X		Y		Z		AA		AB		AC		AD		AE																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																															
																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																										</	



La conversion de données tabulaires semi-structurées en un format typique exploitable par une machine donne le fichier CSV (comma-separated-value en français données séparées par des virgules). Ce sont des fichiers tabulaires avec une ligne d'en-tête qui définit chacune des données dans les colonnes et lignes ci-dessous.



Data.gov dispose d'un [Abécédaire de la lisibilité des documents et données en ligne](https://www.data.gov/developers/blog/primer-machine-readability-online-documents-and-data) (<https://www.data.gov/developers/blog/primer-machine-readability-online-documents-and-data>) et un fil de discussion sur Data.gov.uk propose les conseils suivants :

- En outre, dans la mesure du possible, l'exploitabilité par une machine est améliorée si :



- L'ensemble de données utilise des normes communes lorsqu'elles existent - y compris des identificateurs et des noms de champs standard. Il peut s'agir de normes telles que le vocabulaire des dépenses publiques développé pour le gouvernement, ou de normes tierces telles que KML pour l'indication des "points d'intérêt".

Il est également important de noter qu'il peut être possible de fournir des données dans divers formats exploitables par une machine et, dans la mesure du possible, le fournisseur de données doit échanger avec les réutilisateurs pour identifier le meilleur format. Par exemple, les points d'intérêt pourraient être fournis sous forme de feuille de calcul CSV ou KML. L'idéal serait que les deux soient fournis : toutefois, selon le contexte et le réutilisateur, l'un peut être plus approprié que l'autre.

### 1.3.2 Données *longues* et données *larges*

Toute série de données est constituée de valeurs numériques (généralement) décrites par des termes de métadonnées normalisés (temps, zone, description spécifique, etc.). Il y a deux façons principales de présenter ces données lisibles par machine, qui peuvent être résumées en *largeur* ou *longueur*. Vous devez faire un choix délibéré quant au format que vous choisirez, et chacun a ses propres forces et faiblesses :

- **Les données larges** présentent des données numériques en plusieurs colonnes. Soit sous forme de catégories (par exemple, chaque pays est présenté dans sa propre colonne), soit par date (par exemple, chaque mise à jour annuelle donne lieu à une nouvelle colonne). Les nouvelles données traversent l'écran de gauche à droite :

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

**Les données larges** sont souvent utilisés pour la visualisation et le traitement des données, car les données peuvent facilement être regroupées dans les axes nécessaires aux bibliothèques de cartes. Cependant, il s'agit d'un format d'archivage difficile car la mise à jour d'une telle série de données nécessite l'équivalent de la création d'un nouveau champ (le *année* dans les champs ci-dessus) et ensuite la mise à jour de chaque ligne avec les informations appropriées. Cela peut être une opération coûteuse dans une grande base de données, et signifie également que l'écriture d'une méthode programmatique pour interroger vos données est plus difficile.

- **Les données longues** présente les données numériques sur plusieurs lignes avec une seule colonne pour les valeurs. Les nouvelles données descendent l'écran de haut en bas :

Departme	Entity	Date	Expense T	Expense A	Supplier	Transactio	Amount
Departme	Departme	#####	CASH FUN	FINANCE	NOTTINGH	HAFS-796	45000000
Departme	Departme	#####	CASH FUN	FINANCE	NOTTINGH	HAFS-796	85000000
Departme	Departme	#####	CASH FUN	FINANCE	OLDHAM	HAFS-796	28000000
Departme	Departme	#####	CASH FUN	FINANCE	OXFORDS	HAFS-797	75000000
Departme	Departme	#####	CASH FUN	FINANCE	PETERBOR	HAFS-797	22000000
Departme	Departme	#####	CASH FUN	FINANCE	PLYMOUTH	HAFS-797	35000000
Departme	Departme	#####	CASH FUN	FINANCE	PORTSMO	HAFS-797	27000000
Departme	Departme	#####	CASH FUN	FINANCE	REDBRIDG	HAFS-797	31000000
Departme	Departme	#####	CASH FUN	FINANCE	REDCAR A	HAFS-797	18000000
Departme	Departme	#####	CASH FUN	FINANCE	RICHMON	HAFS-797	21000000
Departme	Departme	#####	CASH FUN	FINANCE	ROTHERH	HAFS-797	29500000
Departme	Departme	#####	CASH FUN	FINANCE	SALFORD	HAFS-797	31000000
Departme	Departme	#####	CASH FUN	FINANCE	SANDWEL	HAFS-797	43000000
Departme	Departme	#####	CASH FUN	FINANCE	SEFTON P	HAFS-798	38000000
Departme	Departme	#####	CASH FUN	FINANCE	SHEFFIELD	HAFS-798	70500000
Departme	Departme	#####	CASH FUN	FINANCE	SHROPSH	HAFS-798	32400000
Departme	Departme	#####	CASH FUN	FINANCE	SOLIHULL	HAFS-798	26200000
Departme	Departme	#####	CASH FUN	FINANCE	SOMERSE	HAFS-798	65000000
Departme	Departme	#####	CASH FUN	FINANCE	SOUTH BIF	HAFS-798	49500000

**Les données longues** sont les meilleures pour l'archivage et pour représenter la structure que vous trouverez habituellement dans une base de données. Chaque ligne d'une *longue* série de données représente une ligne dans une base de données. L'ajout de nouvelles informations est relativement simple puisque vous ne devez mettre à jour qu'une seule ligne à la fois. En termes de base de données, vous créez une seule entrée de base de données.

La préférence dans la publication de données ouvertes est pour le format **long**, et ce sera la méthode habituellement recommandée pour la publication. Cela dit, la conversion entre ces deux formats - pour autant que les données soient exploitables par une machine avec des métadonnées bien définies - est simple.

### 1.3.3 Définir les métadonnées et le schéma de destination

La création d'un schéma est la première partie du processus de préparation des données. Votre schéma définit la structure cible de métadonnées pour votre processus de préparation. Ce n'est pas le format dans lequel vos données d'entrée arrivent, mais c'est ce à quoi vous voulez qu'elle ressemble quand vous avez terminé.

Votre schéma définit les exigences, les contraintes et les valeurs par défaut raisonnables disponibles pour la saisie de données dans les champs définis par le schéma. Une fois terminé, vous devrez procéder à un nettoyage et à une validation supplémentaires.

En termes simples, les colonnes d'un fichier CSV ou Excel d'entrée seront restructurées en de nouvelles colonnes définies par les champs de votre schéma. Ces champs cibles sont

susceptibles d'être ceux d'une base de données. Tant que vos données d'entrée ne sont pas conformes à cette structure, vos données ne doivent pas être publiées en tant que données ouvertes.

Nous utiliserons les définitions et le schéma de [Frictionless Data](https://specs.frictionlessdata.io/table-schema/#field-descriptors) (<https://specs.frictionlessdata.io/table-schema/#field-descriptors>) pour définir les champs de métadonnées structurales.

### ***type et format***

Le `type` définit le type de données du champ, tandis que le `format` affine les propriétés spécifiques du type. Ce sont les types de base que vous êtes susceptible d'utiliser, avec des indentations pour les formats :

- `string` : toute chaîne basée sur du texte.
  - `default` : toute chaîne de caractères
  - `email` : une adresse électronique
  - `uri` : toute adresse web / URI
- `number` : toute valeur basée sur un nombre, y compris les nombres entiers et les nombres flottants.
- `integer` : toute valeur basée sur un nombre entier.
- `boolean` : une valeur booléenne [`true`, `false`]. Peut définir des contraintes de catégorie pour fixer le terme utilisé.
- `object` : toute donnée JSON valide.
- `array` : toute donnée valide basée sur un tableau.
- `date` : toute date sans heure. Doit être au format ISO8601, `YYYY-MM-DD`.
- `datetime` : toute date avec une heure. Doit être au format ISO8601, avec l'heure UTC spécifiée (en option) comme `AAAA-MM-JJ hh:mm:ss Zz`.
- `year` : toute année, formatée comme `YYYY`.

### ***schéma design***

Consultez les données ouvertes présentées sur le site [Données hospitalières relatives à l'épidémie de COVID-19](https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/) (<https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/>). Prévisualisez le fichier de schéma des métadonnées pour l'hospitalisation :

Colonne ↑↓	Type	Description_FR ↑↓	Description_EN ↑↓	Exemple ↑↓
dep	integer	Département	Department	1
sexe	integer	Sexe	Sex	0
jour	string(\$date)	Date de notification	Date of notice	18/03/2020
hosp	integer	Nombre de personnes actuellement hospitalisées	Number of people currently hospitalized	2
rea	integer	Nombre de personnes actuellement en réanimation ou soins intensifs	Number of people currently in resuscitation or critical care	0
rad	integer	Nombre cumulé de personnes retournées à domicile	Total amount of patient that returned home	1
dc	integer	Nombre cumulé de personnes décédées à l'hôpital	Total amount of deaths at the hospital	0

Le premier terme de chaque ligne définit le nom de la colonne de destination (les termes des



métadonnées structurées), ainsi que son type de données, les descriptions en anglais et en français et un exemple de valeur attendue pour ce champ.

Avant de commencer à restructurer des fichiers de données désordonnés, créez une telle définition de métadonnées structurées pour vos données. Et la première chose à faire lorsque vous commencez est simplement de regarder vos données et de comprendre comment elles sont actuellement structurées. L'objectif que vous poursuivez en passant de données désordonnées à des données structurées est simple :

Préserver tout ce qui est connu de vos données sources lorsque vous les restructurez.

Voici un résumé de ces exigences en matière de conception de schémas :

- Etablir une convention standard pour nommer les champs de colonnes, et s'y tenir ; la convention se réfère à la structure physique du mot, comme `startDate` ou `place_name`. Une convention est `casCamel`, une autre est `séparation_substantielle`. Ne les mélangez pas car c'est frustrant pour les utilisateurs.
- Chaque valeur de chaque colonne de chaque table ne doit représenter qu'une seule chose ; si vous regardez la colonne `région` ci-dessous, vous verrez qu'elle doit être divisée en deux colonnes séparées, `région` et `district`, pour la Côte d'Ivoire :

description	région	année	valeur
Population	Gbôklé, Bas-Sassandra	2014	400,798
Population	Nawa, Bas-Sassandra	2014	1,053,084
Population	San-Pédro, Bas-Sassandra	2014	826,666
Population	Indénié-Djuablin, Comoé	2014	560,432
Population	Sud-Comoé, Comoé	2014	642,620

- Définissez le type, le format et la définition pour chaque nom de colonne ; la définition du schéma ci-dessous est la forme **longues**. Si chaque année avait sa propre colonne, vous auriez alors - au lieu de `année` - une ligne pour 2014, alors la prochaine version pourrait inclure 2016 et ainsi de suite. Vous pouvez voir que la mise à jour simultanée des données et du schéma représente plus de travail pour vous et vos utilisateurs que la simple mise à jour des données :

colonne	type	description	exemple
description	string	Définition de la série de données démographiques	Population
région	string	Nom d'une des 31 régions de Côte d'Ivoire	Sud-Comoé
district	string	Nom d'un des 14 districts de la Côte d'Ivoire	Comoé
année	date	L'année de la série de données	2014
valeur	integer	La valeur pour la série de donnée	560432

Il existe des exigences supplémentaires pour valider vos données - par exemple, noter que les virgules entre les chiffres doivent être supprimées - et nous allons les passer en revue ci-dessous.

### **Conception de métadonnées descriptives**

Avant de commencer à débattre de nos données, nous devrions d'abord examiner le tableau suivant des termes formels qui décrivent les données :

OBLIGATOIRE	RECOMMANDÉ	FACULTATIF
Titre	Tag(s)	Dernière mise à jour
Description	Conditions d'utilisation	Fréquence de mise à jour
Thème(s)	Courriel de contact	Couverture géographique
Organisme de publication		Couverture temporelle
		Validité
		Ressources connexes
		Règlements

Ce sont les définitions et les termes de haut niveau pour décrire les données que vous restructurez. Une personne qui ouvre une feuille de calcul a très peu de contexte quant à ce qu'elle lit, et ces métadonnées fournissent un contexte bien nécessaire.

La prochaine étape consiste à commencer à restructurer vos données désordonnées pour les rendre conformes au schéma que vous venez de créer.

### **1.3.4 Les habitudes qui rendent les feuilles de calcul inutilisables**

Dans son excellent essai [The Art of Spreadsheets]

([http://john.raffensperger.org/ArtOfTheSpreadsheet/Chapter09\\_ShowAllTheInformation.html](http://john.raffensperger.org/ArtOfTheSpreadsheet/Chapter09_ShowAllTheInformation.html))

([http://john.raffensperger.org/ArtOfTheSpreadsheet/Chapter09\\_ShowAllTheInformation.html](http://john.raffensperger.org/ArtOfTheSpreadsheet/Chapter09_ShowAllTheInformation.html)))

John Raffensperger énumère 37 façons de cacher des données dans un tableur. En voici 10 :

- Ne pas partager le fichier. C'est la façon la plus courante de cacher des informations, et la plus efficace.
- Cacher la feuille. Il vous faut donc d'abord au moins deux feuilles : Format, Feuille, Cacher.
- Cachez la ligne : Format, Rangée, Cacher.
- Cachez la colonne : Formater, Colonne, Cacher.
- Cacher la cellule et protéger la feuille : Format, Cellules, Protection, Caché, puis Outils, Protection. Ceci affiche un écran, mais cache la formule : =if(1, "Paix !", "Attaque à l'aube.").
- Rendre la colonne trop étroite : Format, Colonne, Largeur, 0.
- Pour les formules susceptibles d'être nulles, utilisez Outils, Options, Affichage et

- Sur les formules susceptibles d'être fautes, amorcez celle, optez, l'ajoutage et désactivez la case des valeurs nulles. Par exemple : =IF(1, 0, "Attaque à l'aube.").
- Utilisez une formule qui renvoie un blanc : =IF(1, " ", "Attaque à l'aube.").
- Créez une formule compliquée qui affiche les informations, mais formatez-les sous forme de texte (avec Format, Cellules, Nombre, Texte, ou commencez simplement la cellule par un simple guillemet), de sorte que la formule soit affichée plutôt que la sortie.
- Formatez la police avec Wingdings : Format, Cells, Font, Wingdings. Cela permet d'afficher des caractères inintelligibles.

### Exercice:

Santé publique France (<https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/>) produit des mises à jour quotidiennes des données d'hospitalisation COVID-19 de la France. En vous inspirant de la liste de John Raffensperger, votre tâche consiste à brouiller autant que possible les données [donnees-hospitalieres-covid19.csv](https://www.data.gouv.fr/fr/datasets/r/63352e38-d353-4b54-bfd1-f1b3ee1cabd7) (<https://www.data.gouv.fr/fr/datasets/r/63352e38-d353-4b54-bfd1-f1b3ee1cabd7>).

Les notes seront attribuées pour:

- en rendant la présentation juste assez mauvaise pour que quelqu'un qui utilise les données soit tenté de penser qu'il peut encore les utiliser !
- l'utilisation d'effets de couleurs et de polices de caractères d'une manière qui offense vraiment l'œil
- ingéniosité pour cacher des données à la vue de tous.

## 1.3.5 Utiliser Excel pour nettoyer les données désordonnées

Excel est l'un des logiciels les plus puissants utilisés par les analystes ordinaires (non développeurs de logiciels), et probablement l'outil de gestion et d'analyse de données le plus répandu dans le monde.

Excel peut vous permettre de faire tout ce qui suit :

- Supprimer les enregistrements en double
- Séparer les valeurs multiples contenues dans le même champ
- Analyser la distribution des valeurs dans un ensemble de données
- Regrouper les différentes représentations d'une même réalité

Un guide simple, étape par étape, pour la préparation et le nettoyage des données :

- **Créer le fichier de données** avec de nouvelles feuilles de travail pour chacune d'entre elles : Données originales | Données intermédiaires | Données finales
- **Nettoyer les données :**
  - Créez une colonne ID reliant vos OriginalData à vos InterimData pour garder la trace des lignes supprimées ;
  - Gérer les enregistrements en double en créant une clé de recherche qui doit identifier chaque ligne de manière unique : Excel peut trier sur cette ligne et

identifier chaque ligne de manière unique ; Excel peut agir sur cette ligne et supprimer les doublons ;

- Supprimer les caractères indésirables (rechercher/remplacer) ;
- Repérer les valeurs hors limites (c'est-à-dire les valeurs qui ne sont pas ce qu'elles devraient être) ;
- Supprimer toute donnée non publiable (comme les informations personnelles ou toute donnée non approuvée pour la publication) ;

- **Traiter les données :**

- analyser les données (par exemple, en utilisant la méthode du texte en colonnes et en les séparant par des espaces ou des tabulations) ;
- Recoder les données (pour convertir les termes en métadonnées normalisées) ;
- Calculer de nouvelles valeurs (comme des totaux ou des moyennes) ;
- Reformater les données dans un format standardisé ;

- **Créer une copie des données prête** pour l'analyse en copiant les InterimData non formatées dans les FinalData et en supprimant les colonnes inutiles ;

- **Documenter les données :**

- Documentation au niveau du fichier, telle que le projet, la date d'achèvement, vérifiée par ;
- Produire un fichier texte qui accompagne le fichier de données et qui comporte : description, date de production, propriétaire des données, etc. (à partir des métadonnées décrites dans l'étape de classification et de traitement) ;

### Tutoriels et références:

- Quand les données lisibles par machine posent encore des "problèmes" - Les dates de dispute... (<http://blog.ouseful.info/2012/11/27/when-machine-readable-data-still-causes-issues-wrangling-dates/>)
- Paysages numériques : des données ouvertes efficaces nécessitent plus qu'un simple CSV (<http://www.opendataimpacts.net/2012/11/more-than-csv/>)
- Comment préparer vos données pour l'analyse et les graphiques dans Excel & Google Sheets (<https://blog.datawrapper.de/prepare-and-clean-up-data-for-data-visualization/>)
- Top dix façons de nettoyer vos données (<https://support.office.com/en-us/article/Top-ten-ways-to-clean-your-data-2844b620-677c-47a7-ac3e-c2e157d1db19>)
- Un guide pour la conception de schémas de bases de données (<https://www.mikealche.com/software-development/a-humble-guide-to-database-schema-design>)

## 1.4 Exemple de préparation de données

L'exemple suivant a été écrit par Lisa Charlotte Rost @ Datawrapper.de



(<https://blog.datawrapper.de/prepare-and-clean-up-data-for-data-visualization/>) et est utilisée avec son autorisation. Il a été légèrement modifié pour être conforme à l'approche adoptée dans ce cours. Il s'agit d'un exemple relativement simple, mais qui vous guidera à travers les principaux défis que vous rencontrerez.

Le tutoriel travaillé utilise des données de la [Banque mondiale] (<https://data.worldbank.org/indicator/SP.URB.TOTL?view=chart>) (<https://data.worldbank.org/indicator/SP.URB.TOTL?view=chart>). Un lien vers les métadonnées ouvre la fenêtre suivante :

×

## Urban population

Urban population refers to people living in urban areas as defined by national statistical offices. It is calculated using World Bank population estimates and urban ratios from the United Nations World Urbanization Prospects. Aggregation of urban and rural population may not add up to total population because of different country coverages.

**ID:** SP.URB.TOTL

**Source:** World Bank staff estimates based on the United Nations Population Division's World Urbanization Prospects: 2018 Revision.

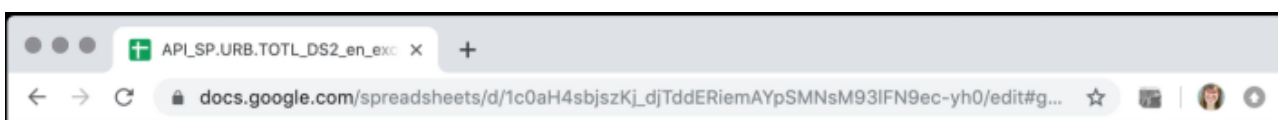
**License:** CC BY-4.0 [↗](#)

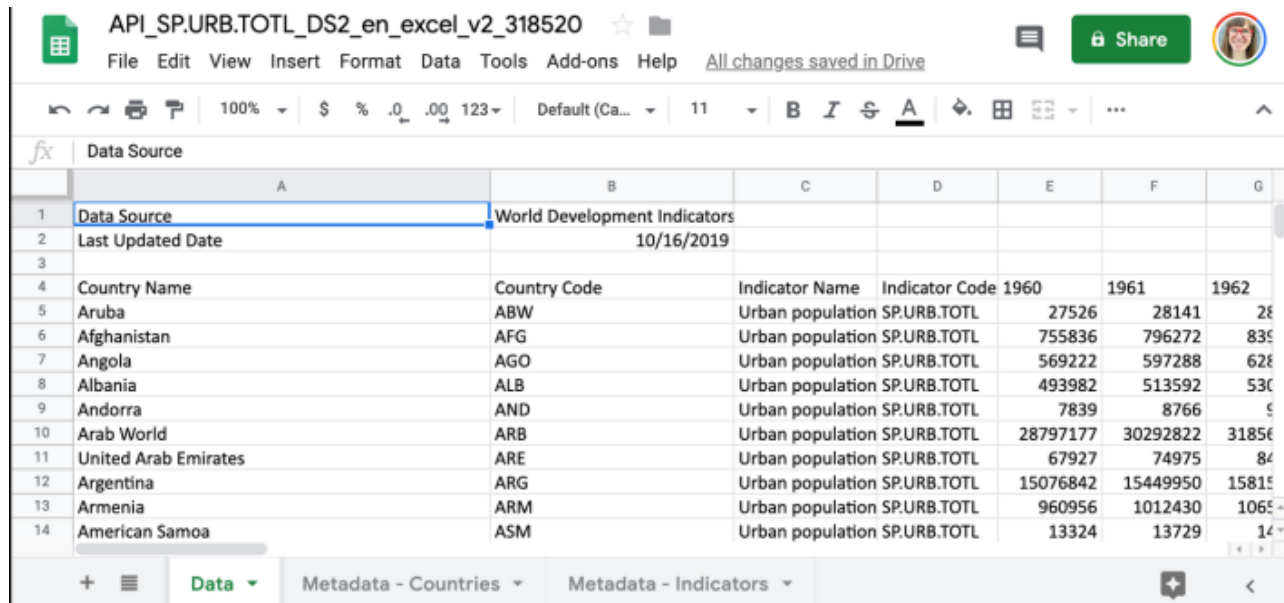
**Aggregation Method:** Sum

**Development Relevance:** Explosive growth of cities globally signifies the demographic transition from rural to urban, and is associated with shifts from an agriculture-based economy to mass industry, technology, and service. In principle, cities offer a more favorable setting for the resolution of social and environmental problems than rural areas. Cities generate jobs and income, and deliver education, health care and other services. Cities also present opportunities for social mobilization and women's empowerment.

### 1.4.1 Examiner les données

Téléchargez cet ensemble de données (<http://api.worldbank.org/v2/en/indicator/SP.URB.TOTL?downloadformat=excel>) et consultez la fiche principale.





	A	B	C	D	E	F	G
1	Data Source	World Development Indicators					
2	Last Updated Date	10/16/2019					
3							
4	Country Name	Country Code	Indicator Name	Indicator Code	1960	1961	1962
5	Aruba	ABW	Urban population SP.URB.TOTL		27526	28141	28141
6	Afghanistan	AFG	Urban population SP.URB.TOTL		755836	796272	835
7	Angola	AGO	Urban population SP.URB.TOTL		569222	597288	628
8	Albania	ALB	Urban population SP.URB.TOTL		493982	513592	530
9	Andorra	AND	Urban population SP.URB.TOTL		7839	8766	9
10	Arab World	ARB	Urban population SP.URB.TOTL		28797177	30292822	31856
11	United Arab Emirates	ARE	Urban population SP.URB.TOTL		67927	74975	84
12	Argentina	ARG	Urban population SP.URB.TOTL		15076842	15449950	15815
13	Armenia	ARM	Urban population SP.URB.TOTL		960956	1012430	1065
14	American Samoa	ASM	Urban population SP.URB.TOTL		13324	13729	14

Lorsque vous téléchargez un fichier Excel, il comporte souvent plusieurs feuilles. Notre ensemble de données en comporte trois, comme on peut le voir en bas : "Données", "Métadonnées - Pays" et "Métadonnées - Indicateurs". Regardez toutes vos feuilles et assurez-vous que vous comprenez ce que vous y voyez. Les en-têtes, le nom du fichier et/ou les données elles-mêmes indiquent-ils que vous avez téléchargé le bon fichier ? Y a-t-il des notes de bas de page ? Que vous disent-elles ? Que vous avez peut-être affaire à de nombreuses estimations ? (Cela signifie-t-il peut-être que vous devez chercher d'autres données ?) Si vous ne trouvez pas de notes dans les données, assurez-vous de les chercher sur le site web de votre source.

Notre exemple de données semble correct. Il n'y a pas d'estimations dont nous devrions nous préoccuper. Et nous obtenons une belle explication de la "population urbaine" dans les "Métadonnées - Indicateurs", qui commence par "La population urbaine se réfère aux personnes vivant dans des zones urbaines telles que définies par les bureaux nationaux de statistiques...". Génial ! C'est quelque chose que nous pouvons mentionner dans notre tableau plus tard.

## 1.4.2 Renommer votre fichier

Maintenant que nous savons à quoi nous avons affaire, assurons-nous de le faire encore dans six mois. "API\_SP.URB.TOTL\_DS2\_fr\_excel\_v2\_318520 ! Oui ! Je sais exactement de quoi il s'agit": vous ne trouverez personne qui dira jamais une chose pareille sauf peut-être trois employés en tout à la Banque mondiale.) Alors renommons ce fichier en quelque chose de mémorisable et de précis : Banque mondiale\_population\_urbaine\_par-pays, par exemple.

## 1.4.3 Dupliquer la ou les fiches de données et ne plus jamais y toucher

C'est l'une des parties les plus importantes de l'ensemble du processus : Avant de modifier quoi que ce soit dans les données, dupliquez votre feuille de données dans le même fichier.

Pensez à renommer vos deux feuilles. par exemple en "Données brutes" et "Données" ou en

... et renommer les deux feuilles, par exemple en "Données brutes" et "Données".

"Données - original" et "Données - modifié". Si vous avez un énorme fichier Excel avec beaucoup de feuilles, vous pouvez également créer un fichier de sauvegarde du document original et le stocker dans un dossier "source".

Pourquoi devriez-vous faire cela ? Parce que vous allez modifier considérablement les données. J'ai appris à mes dépens que je modifie toujours les données plus que je ne n'anticipe!. "Je n'ai pas besoin de copier les données cette fois-ci." Je pense. "Je veux seulement les nettoyer un peu ; je ne supprimerai rien d'important." Deux heures passent... et j'ai besoin de télécharger à nouveau les données à partir de leur source originale parce que, oh oui, j'ai supprimé cette colonne maintenant importante il y a une heure. Apprenez de mes erreurs, gagnez beaucoup de temps et ne modifiez jamais les données originales.

#### 1.4.4 Enregistrez votre source dans une feuille supplémentaire

Cette astuce vous donnera également envie de vous taper sur l'épaule et de dire "Merci" : Créez une nouvelle feuille, nommez-la "Source" et ajoutez des liens vers toutes les sources de données que vous utilisez dans votre document. (Et oui, vous obtenez des points bonus pour avoir ajouté la date de téléchargement du fichier - au cas où).

#### 1.4.5 Supprimer tout ce qui se trouve au-dessus de l'en-tête

Les fichiers Excel contiennent souvent des informations dans des lignes supplémentaires au-dessus des données réelles. Dans notre cas, de sympathiques employés de la Banque mondiale veulent nous faire savoir que la source des données est les "Indicateurs du développement dans le monde", et que la dernière mise à jour des données remonte à octobre 2019. C'est à la fois bon à savoir et quelque chose que nous pouvons mettre dans le tableau. Mais ces lignes supplémentaires nous empêchent de trier ou de filtrer les données.

Il suffit de supprimer toutes les lignes vides et toutes les informations au-dessus de l'en-tête. Supprimez-les (vous pouvez toujours vérifier votre feuille de "données brutes" lorsque vous avez besoin de ces informations) ou copiez-collez les informations dans votre feuille "Source".

	A	B
1	Data Source	World Development Indicat
2	Last Updated Date	10/16/20
3		
4	Country Name	Country Code
5	Aruba	ABW
6	Afghanistan	AFG
7	Angola	AGO

#### 1.4.6 Séparer les cellules fusionnées et enlever les entêtes sur deux ou plusieurs lignes

## ou plusieurs lignes

Parfois, vous rencontrerez des en-têtes qui sont sur deux lignes, et non une seule. Surtout lorsque le tableau est créé pour communiquer, et non pour analyser, les en-têtes à deux lignes peuvent aider à donner un sens à l'information. Mais ils vous gêneront lorsque vous voudrez éventuellement supprimer des lignes ou des colonnes. Les données exploitables par une machine doivent comporter une ligne d'en-tête, et une seule ligne d'en-tête.

Démêlez les lignes doubles avec des cellules fusionnées comme ceci :

Afghanistan		Angola	
Population	Urban population	Population	Urban population

A cela :

Afghanistan Population	Afgahnistan Urban population	Angola Population	Angola Urban population
---------------------------	---------------------------------	----------------------	----------------------------

Il en va de même pour les cellules fusionnées. Peu importe où elles se trouvent dans votre ensemble de données, éliminez-les :

Afghanistan	1960	755836
	1961	796272
	1962	839385
	1963	885228
	1964	934135
	1965	986074

### 1.4.7 Mettre des mesures dans l'en-tête et supprimer les notes de bas de page

Pour que les outils de données tels que Datawrapper et Excel reconnaissent les chiffres, assurez-vous que vous disposez de chiffres non perturbés dans vos cellules de données. Non perturbés par des milliers de séparateurs - mais nous nous occuperons de cela plus tard - et non perturbés par des mesures. Libérez donc vos cellules de données de tous les €, \$, kg, %, km/h, etc. Mettez-les plutôt dans les en-têtes.

Cela nécessitera souvent une série d'étapes de recherche - remplacement, mais assurez-vous toujours d'avoir créé une colonne supplémentaire capturant ces unités afin de ne pas les perdre. Cette opération est similaire à la division des valeurs qui contiennent plusieurs termes.

Cela signifie qu'il faut partir des métriques dans les cellules de données :

	<b>fake GDP</b>	<b>fake share</b>
Afghanistan	293 Euro	46.48%
Angola	383 Euro	29.40%
Albania	919 Euro	19.01%
Andorra	294 Euro	31.07%

Pour cela, il faut les séparer dans leur propre colonne, ou - comme ici - les capturer dans le champ nom de la colonne :

	<b>fake GDP in Euro</b>	<b>fake share in %</b>
Afghanistan	293	46.48
Angola	383	29.40
Albania	919	19.01
Andorra	294	31.07

Une chose à noter dans l'exemple ci-dessus est que le % serait normalement une proportion de 1 (par exemple, 48% est 0,48), et donc les données dans cette colonne sont susceptibles d'être ambiguës. Il convient de corriger cette situation pour garantir la validité de vos données.

Vous devrez également extraire les notes de bas de page. Des valeurs comme 28.394+ ou 1.39 [ ^ 1 ] ne seront pas reconnues comme des chiffres.

Avant de les supprimer, assurez-vous que vous comprenez le schéma des données : Les points de données de 2019 sont-ils tous des estimations ? (Devriez-vous peut-être exclure cette année-là ?) Ou bien les données d'un certain pays sont-elles mesurées différemment ? Dans tous ces cas, veillez à le faire savoir au lecteur de votre tableau final. Les notes de bas de page dans les données que vous utilisez doivent toujours se traduire par des définitions spécifiques dans le fichier de métadonnées. Vos utilisateurs en auront besoin.

### 1.4.8 Vérifier si le contenu de l'en-tête a un sens

Après avoir effectué ces tâches techniques, voyons si les noms de l'en-tête que vous n'avez pas encore touchés ont vraiment un sens. Peut-être ne sont-ils que du charabia de code comme SP.URB.TOTL ? Si c'est le cas, retournez à votre source et découvrez la signification de ces codes. Ou peut-être sont-ils trop longs ? Par exemple, Nom du pays peut facilement être réduit à Pays. Renommez les en-têtes pour qu'il soit facile pour les étrangers de leur donner un sens : courts, mais précis et uniques. (Si vous vovez des colonnes que vous avez

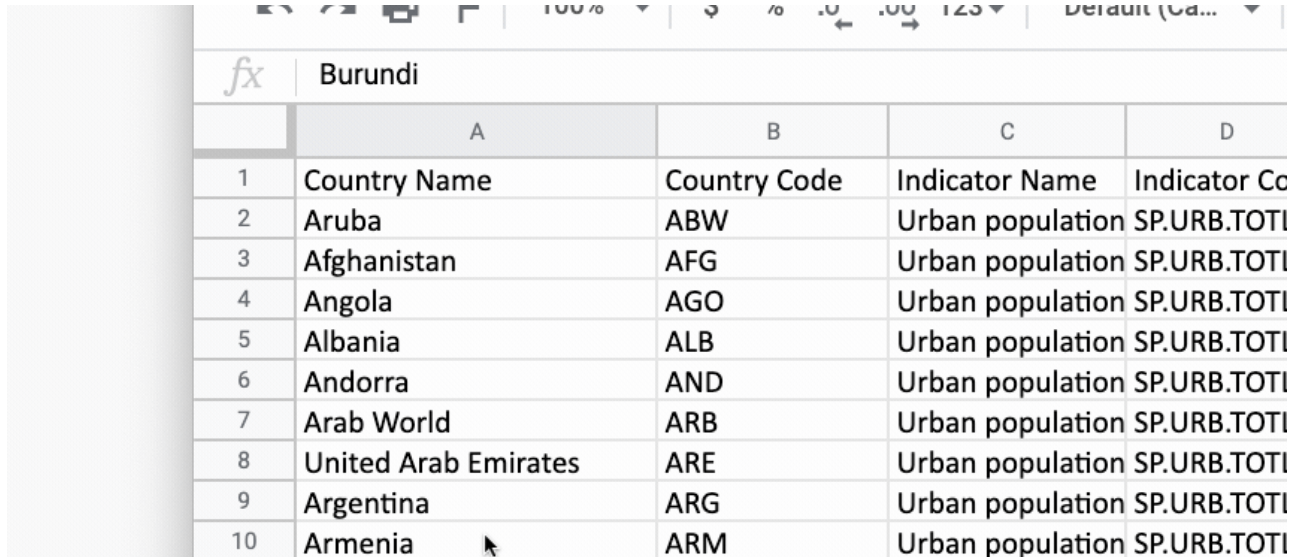


l'intention de supprimer, ne vous embêtez pas à les renommer).

Rappelez-vous les conventions de dénomination dont nous avons parlé plus haut. Vous pouvez aussi choisir de convertir `Nom du pays` en `nom du pays`, mais soyez cohérent dans la façon dont vous nommez tous vos champs.

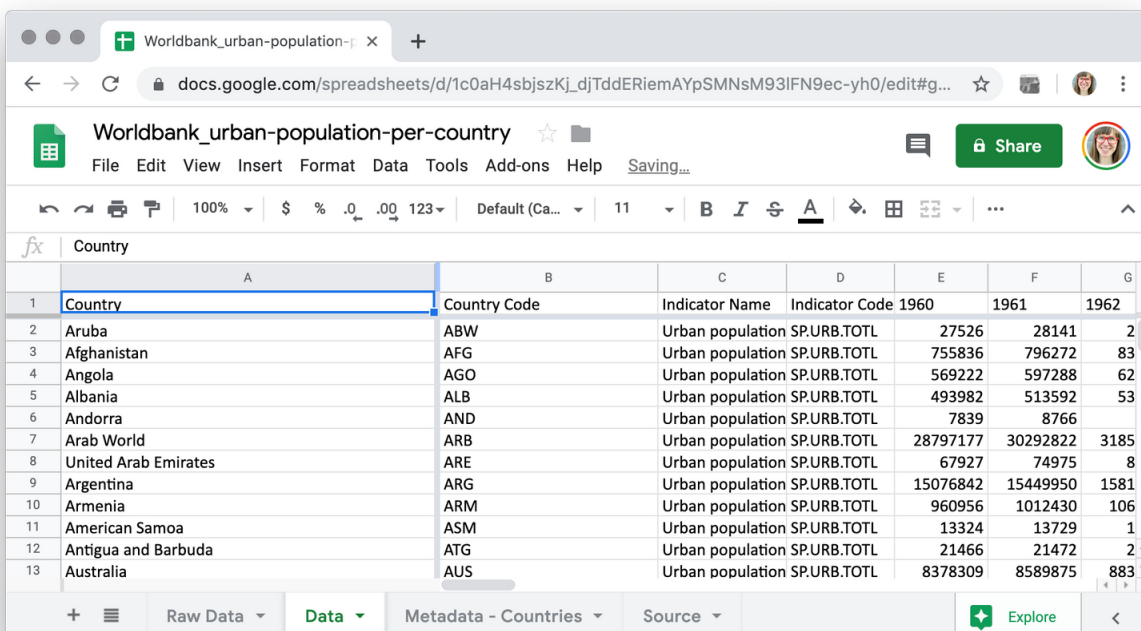
### 1.4.9 Parce que c'est pratique : Geler la première ligne (et la première colonne)

Vous devriez maintenant être à un point où vos en-têtes sont excellents. Félicitations ! Assurez-vous de toujours les avoir beautés en vue et figez les :



	A	B	C	D
1	Country Name	Country Code	Indicator Name	Indicator Code
2	Aruba	ABW	Urban population SP.URB.TOTL	
3	Afghanistan	AFG	Urban population SP.URB.TOTL	
4	Angola	AGO	Urban population SP.URB.TOTL	
5	Albania	ALB	Urban population SP.URB.TOTL	
6	Andorra	AND	Urban population SP.URB.TOTL	
7	Arab World	ARB	Urban population SP.URB.TOTL	
8	United Arab Emirates	ARE	Urban population SP.URB.TOTL	
9	Argentina	ARG	Urban population SP.URB.TOTL	
10	Armenia	ARM	Urban population SP.URB.TOTL	

Voilà à quoi ressemblent nos données maintenant. C'est déjà un peu plus propre :



	A	B	C	D	E	F	G
1	Country	Country Code	Indicator Name	Indicator Code	1960	1961	1962
2	Aruba	ABW	Urban population SP.URB.TOTL		27526	28141	2
3	Afghanistan	AFG	Urban population SP.URB.TOTL		755836	796272	83
4	Angola	AGO	Urban population SP.URB.TOTL		569222	597288	62
5	Albania	ALB	Urban population SP.URB.TOTL		493982	513592	53
6	Andorra	AND	Urban population SP.URB.TOTL		7839	8766	
7	Arab World	ARB	Urban population SP.URB.TOTL		28797177	30292822	3185
8	United Arab Emirates	ARE	Urban population SP.URB.TOTL		67927	74975	8
9	Argentina	ARG	Urban population SP.URB.TOTL		15076842	15449950	1581
10	Armenia	ARM	Urban population SP.URB.TOTL		960956	1012430	106
11	American Samoa	ASM	Urban population SP.URB.TOTL		13324	13729	1
12	Antigua and Barbuda	ATG	Urban population SP.URB.TOTL		21466	21472	2
13	Australia	AUS	Urban population SP.URB.TOTL		8378309	8589875	883

## 1.4.10 Supprimer les colonnes et lignes inutiles

Vos données sont destinées à être des données sources, il devrait donc être rare que vous ayez besoin de supprimer quoi que ce soit, mais parfois les données sont redondantes. Soit parce que les données sont répétitives, soit parce qu'elles fournissent des résumés qui ne sont pas nécessaires parce qu'ils peuvent être calculés.

Par exemple, si vous disposez de données sur les dépenses par poste, avez-vous également besoin d'une ligne pour le `total` ? C'est une décision que vous devrez prendre, mais n'oubliez pas de la consigner dans votre fichier de métadonnées descriptives.

## 1.4.11 Supprimer les milliers de séparateurs

Les milliers de séparateurs sont des caractères ( , en anglais, . en allemand, parfois c'est juste un espace) qui permettent de reconnaître facilement la magnitude d'un nombre. Par exemple, 38.394.105 s'arrondit plus vite à 38 millions dans notre esprit que 38394105.

Mais s'ils sont géniaux et utiles pour les humains, ils sont terribles pour les ordinateurs et conduisent à des interprétations ambiguës. Débarrassons-nous de toute sorte de milliers de séparateurs.

Trouvons et remplaçons des milliers de séparateurs pour partir de là :

755,836	796 272
569,222	597 288
493,982	513 592
7,839	8 766
28,797,177	30 292 822
67,927	74 975

A cela :

--	--