Produire un fichier CSV de qualité

Connaître les incontournables et les mettre en oeuvre

Source : FING (Charles Népote, Armelle Gilliard), juin 2017, v1.1, Licence : Creative Commons 3.0 Attribution France.
http://infolabs.io/prod-csv

Le format CSV est le standard le plus simple et le plus répandu pour échanger des données numériques organisées en tableau. Que ce soit pour un projet open data ou pour tout autre projet d'échange de données, sa connaissance est primordiale.

Pour réaliser un fichier de qualité, ce document propose 3 modes de lecture.

➔ **Pour le lecteur averti**

Aguerri aux données ? Ce simple encart devrait vous suffire.

Un fichier CSV de qualité :

1. est encodé en UTF-8,
2. utilise CRLF pour chaque fin de ligne,
3. utilise la virgule comme délimiteur,
4. encapsule entre guillemets les champs dont le contenu possède une ou plusieurs virgules, par exemple : Jean,Martin,"lundi, mardi, jeudi"
5. possède un en-tête (la première ligne du fichier) où chaque champ est décrit par un libellé,
6. est validé par l'outil en ligne CSVLint : http://csvlint.io/,
7. est documenté.

Mise en oeuvre simple : un tableur, comme Excel ou LibreOffice, peut suffire à produire un fichier de qualité. Les éventuelles erreurs de syntaxe peuvent être traitées à la main à la suite du rapport de CSV Lint.

Aller plus loin et industrialiser : l'outil *csvclean*, du package *csvkit*, permet de nettoyer un fichier de manière automatique (y compris de très gros fichiers) : https://csvkit.readthedocs.io/ (il est détaillé en page 6)

➔ **Pour le lecteur curieux**
Le reste de ce document explique le pourquoi et le comment de la norme ainsi que les bonnes pratiques pour produire un CSV.

➔ **Pour le lecteur néophyte mais pressé**
En annexe (2 dernières pages), nous avons détaillé chaque étape pratique à partir des deux tableurs les plus répandus, Excel et LibreOffice Calc.

Aidez-nous à améliorer ce document !
Nous sommes à l'écoute de vos retours. Racontez-nous vos incompréhensions, vos problèmes, les manques voire les erreurs de ce document : armelle@lareinemerlin.org et charles.nepote@fing.org.

Qu'est-ce que le CSV ? Pourquoi le CSV ?

Le format CSV est le standard le plus simple et le plus répandu pour échanger des données numériques organisées en tableau. Il se présente sous une forme simple à interpréter par un logiciel ou toute autre forme de programme informatique. Mais il est également lisible par un humain car sa forme et sa syntaxe sont rudimentaires : il s'agit d'un fichier texte contenant des valeurs séparées par un caractère spécial — en anglais, CSV signifie *Comma Separated Values* soit littéralement "valeurs séparées par des virgules". Voici comment se présente, dans un éditeur de texte, un fichier CSV simple :

```
Prénom,Nom,Age
Marie,Durand,37
Bernard,Martin,29
```

Dans un tableur, ce fichier donnera le résultat suivant :

	A	B	C
1	Prénom	Nom	Age
2	Marie	Durand	37
3	Bernard	Martin	29

Si votre fichier est correctement produit, il sera lisible sans effort par des logiciels et/ou programmes informatiques usuels : tableurs, logiciels de statistique, logiciels de traitement de données spécialisés, etc. Autrement dit, si vous souhaitez partager correctement vos données, il est important de les publier sous forme d'un fichier CSV de bonne qualité.

La norme CSV
Du fait de sa simplicité, le format CSV est utilisé depuis des temps immémoriaux. À tel point qu'il était très utilisé bien avant sa normalisation, sous de nombreuses variantes (on parle de "dialectes", ces derniers sont encore répandus). En octobre 2005, il est finalement spécifié à travers la RFC 4180 intitulée *Common Format and*

MIME Type for Comma-Separated Values (CSV) Files (https://tools.ietf.org/html/rfc4180).
À l'origine, la RFC 4180 n'avait pas la prétention de devenir un standard : cette dernière évoquait seulement le fait de décrire la forme de CSV la plus courante. Avec le temps, et pour gagner en interopérabilité, la RFC est cependant devenu le standard *de facto*. Les outils dédiés au CSV respectent souvent et de plus en plus la RFC : la suivre est donc un gage de meilleure réutilisabilité des données.

Quelles données ?

Son aspect rudimentaire fait du CSV un format très simple à réutiliser. Ce côté rudimentaire à d'autres conséquences pratiques : il ne mémorise pas les couleurs, les onglets, les cellules fusionnées, les tailles de caractère... Il n'est adapté qu'à des tableaux simples où tous les enregistrements, c'est-à-dire toutes les lignes, ont la même forme. C'est strictement un format de données qui ne peut pas accepter de mise en forme. Certains tableaux, pensés comme des documents autant que des données devront faire l'objet d'une préparation pour être exporté en CSV.
ATTENTION : dans votre tableur préféré, si vous pratiquez des mises en forme de votre fichier CSV, elles seront perdues au moment de l'enregistrement.

Nommer un fichier CSV

Il n'existe aucune norme pour nommer un fichier mais rappelons quelques points de bon sens :

- un nom trop générique, comme "liste.csv", risque d'entraîner des confusions
- un nom trop long sera difficile à manipuler
- un nom contenant des caractères spéciaux ou accentués risque de poser des problèmes d'interopérabilité

L'idéal est de vous fixer deux ou trois règles simples et de vous y tenir. Une bonne pratique consiste à composer ce nom avec une partie qui vous identifie (code INSEE ou SIREN). La présence d'une date peut aider. Par exemple : 34172_Geoloc_ArbresRemarquables_2014.csv renseigne sur le contenu sans avoir à ouvrir le fichier.

L'encodage du fichier

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'US-ASCII, l'ISO-Latin-1 et l'UTF-8 sont les plus répandus en France. 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'encodage 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, comme par exemple l'alphabet grec, le cyrillique, les alphabets japonais, chinois, arabe, hébreux, etc.

L'UTF-8 est en passe de devenir le standard de référence universel. En janvier 2017 le site W3Techs recense que 88,4% des sites web analysés utilisent l'UTF-8.

Pour toutes ces raisons, **un fichier CSV de qualité est donc encodé en UTF-8**.

Le type de fin de ligne

Un fichier CSV est constitué de lignes représentant chacune un enregistrement. Par exemple, le code suivant contient 3 enregistrements (la première ligne, l'en-tête, n'est pas comptée) :

```
"Prénom","Nom","Note"
"Marie","Durand","13,4"
"Bernard","Martin","12"
"Célestin","Lampion","9"
```

Chaque ligne est séparée de la précédente par un ou plusieurs caractères invisibles : la fin de ligne. Les différents systèmes d'exploitation (Windows, Mac OS, Linux) utilisent cependant un code différent pour la fin de ligne. La norme préconise l'emploi de la combinaison [CR]+[LF] correspondant à l'usage standard sous Windows. De nos jours, la plupart des outils savent traiter des fichiers quelque soit leur caractère de fin de ligne. Ce n'est cependant pas le cas du programme "Notepad", utilisé par défaut sous Windows pour ouvrir les fichiers .txt ou les fichiers .csv.
L'usage de **la combinaison [CR]+[LF] reste donc préférable dans tous les cas** pour maximiser le potentiel de réutilisation des données.

Détecter et modifier l'encodage et les fins de lignes de mon fichier

Il existe de très nombreuses façons de faire. Une des plus simples consiste à utiliser un éditeur de texte. Les éditeurs suivants (logiciels libres, non limitatif) gèrent plutôt bien les caractères de fin de ligne et l'encodage :

- Notepad++ (Windows) ; https://notepad-plus-plus.org/
- Geany (Windows, Mac OSX, Linux) ; http://www.geany.org/
- Atom (Windows, Mac OSX, Linux), en installant le plug-in adéquat ; https://atom.io/

Geany, en particulier, affiche l'encodage en bas de l'écran et possède un menu spécial pour changer l'encodage et les caractères de fin de ligne. Voici les étapes nécessaires :

- Ouvrir le fichier concerné avec Geany.
- Observer au bas de la fenêtre le "mode" de fins de ligne et le "codage" du fichier
- Si le "mode" et le "codage" ne correspondent pas à "CRLF" et "UTF-8" :
  - Menu Document > Définir les fins de lignes > Convertir en CRLF (Windows)
  - Menu Document > Définir l'encodage > Unicode > Unicode (UTF-8)
  - Menu Fichier > Enregistrer. Et voilà !

Le séparateur utilisé

Le séparateur, ou délimiteur, est le caractère qui permet à un programme de distinguer les cellules les unes des autres. Dans le cas suivant le séparateur est la virgule :

```
Prénom,Nom,Age
Marie,Durand,37
Bernard,Martin,29
```

Le séparateur est aussi fréquemment un point-virgule, une tabulation [□] ou le caractère | (dit *barre verticale* ou *tube* en français, ou "pipe" en anglais[1]). Le délimiteur peut encore être n'importe quel caractère du moment qu'il permette de séparer les champs sans ambiguïté.
Cependant, **la norme CSV désigne la virgule comme le caractère à utiliser**.

Cet usage peut nous poser problème à nous autres français, car cette dernière est notamment utilisée comme séparateur décimal… Un "3,5" caché au milieu de nombres entiers pourra passer inaperçu et provoquera des erreurs de lecture.
Pour autant de nombreux outils de traitement du format CSV attendent par défaut l'usage de la virgule. **Le séparateur idéal reste donc la virgule** mais il faut alors **encapsuler chaque champ entre des guillemets** (au moins ceux qui contiennent une virgule) :

```
"Prénom","Nom","Note"
"Martin","Durand","13,4"
```

Pas de panique ! en utilisant un tableur comme Excel ou OpenOffice Calc, ces derniers réaliseront automatiquement l'encapsulation des champs entre guillemets.

[1] L'article Wikipedia est bien documenté et indique comment le produire selon les différents clavier : *Barre verticale*, https://fr.wikipedia.org/wiki/Barre_verticale

L'en-tête de description des champs

La première ligne peut être utilisée pour nommer chaque colonne, on l'appelle alors l'en-tête (comme dans l'exemple ci-contre).

```
Prénom,Nom,Age
Marie,Durand,37
Bernard,Martin,42
```

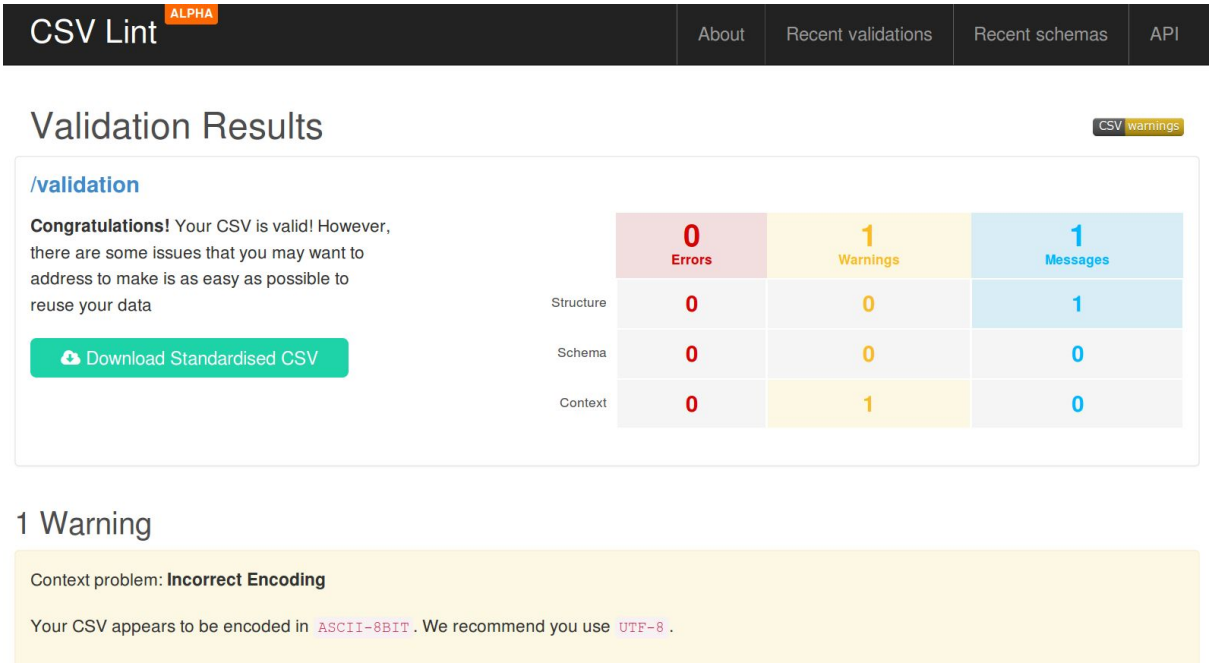
L'en-tête n'est pas obligatoire mais il **augmente sensiblement la qualité** du jeu de données puisqu'il permet d'identifier chaque colonne et donc de lever d'éventuelles ambiguïtés. Il est plus approprié de nommer une colonne "date_naissance" que de la désigner par "la quatrième colonne" cela permet au lecteur de comprendre rapidement le sens du champ concerné. De plus, au fur et à mesure de l'évolution du jeu de données, "la quatrième colonne" pourrait se trouver à une autre place. Bien nommer les colonnes rend le fichier et sa documentation plus compréhensibles et faciles à utiliser, comme le fait un index.
Enfin, ce procédé est très facile à mettre en oeuvre. Il est dommage de s'en priver !

Contrôle syntaxique minimal

Il est très difficile de contrôler à la main des centaines de lignes d'un fichier CSV. Une erreur, quelle qu'elle soit, a pu se glisser dans les étapes successives de production des données. Un contrôle syntaxique automatique doit permettre de garantir qu'un fichier pourra être exploité par n'importe quel outil, logiciel ou programme informatique.
Pour ce faire, nous présentons ici CSV Lint, solution simple à mettre en oeuvre. Nous proposons également plus bas une autre solution pour les utilisateurs avancés, csvclean.

CSV Lint

CSV Lint est un service en ligne — hélas anglophone — qui établit un rapport signalant les problèmes élémentaires de votre jeu de données (encodage, délimiteur, syntaxe, etc.). Son usage est relativement simple. Il suffit de téléverser (*uploader*) le fichier CSV désiré et lancer la validation : l'outil affiche alors un rapport d'analyse complet signalant tous les problèmes identifiés (voir copie d'écran ci-dessous).



Pour les utilisateurs avancés : *csvclean*

csvclean est un outil pour utilisateurs avancés qui savent manier la ligne de commande. Il est plus frustre à utiliser mais permet de traiter des fichiers de TRÈS grande taille (testé avec succès sur le CSV de la base SIRENE de plus de 8 Go !).
Le logiciel fonctionne sous Windows, Mac OS X et Unix/Linux. Nous renvoyons à la documentation de l'outil pour son installation : https://csvkit.readthedocs.io

L'usage de *csvclean* est extrêmement simple. Pour contrôler le fichier :

```
$ csvclean fichier.csv --dry-run
```

Si nécessaire les options -e et -d spécifient l'encodage du fichier et le délimiteur :

```
$ csvclean -e iso-8859-1 -d ";" fichier.csv --dry-run
```

csvclean va au-delà du contrôle et permet de corriger le fichier :

```
$ csvclean fichier.csv
```

À l'heure où nous écrivons ces lignes (14/02/2017), *csvclean* possède la fâcheuse "fonctionnalité" de convertir les fins de ligne en [LF] et non [CR]+[LF] comme le préconise la norme. Une fois le traitement de *csvclean* effectué, il faut donc convertir les fins de ligne, à l'aide de l'outil csvformat, issu du même "toolkit" :

```
$ csvformat -M $'\r\n' fichier_out.csv > tmp && mv tmp fichier_out.csv
```

Documenter un fichier CSV

Sans documentation, un jeu de données quel qu'il soit, est très compliqué à utiliser. Les futurs utilisateurs ont besoin de comprendre ce qu'il contient, à quoi correspondent les différentes composantes du fichier, comment les données ont été collectées, etc.

Document spécifique publié à part, il devrait idéalement contenir les rubriques suivantes.

- Le titre du jeu de données.
- La description du jeu : il s'agit de quelques paragraphes décrivant le jeu de données : son usage, son contexte et son mode de production, ses producteurs.
- La fréquence de mise à jour des données.
- La date de création du jeu.
- La date de dernière mise à jour du jeu.
- La couverture temporelle (ex. année 2014, période 2009-2016...).
- La granularité de la couverture temporelle (tous les 2 ans, annuelle, trimestrielle…)
- La granularité de la couverture territoriale.
- La license appliquée au jeu de données.
- La description de chaque champ du jeu de données :
    - le libellé du champ (le nom que vous avez retenu pour la colonne)
    - sa position (colonne 1 ou colonne 2 ou …)
    - sa description "sémantique" : que signifie-t-il ?
    - mais aussi sa syntaxe : sa longueur maximum, son type de contenu (chaîne de caractère ? nombre ? booléen ? date ?...), son format

Comme vous pourrez le constater dans l'exemple suivant, il est facile de bien documenter : http://www.hatvp.fr/consulter-les-declarations/#open-data

En guise de conclusion pour aller plus loin : la norme de description *Table Schema*

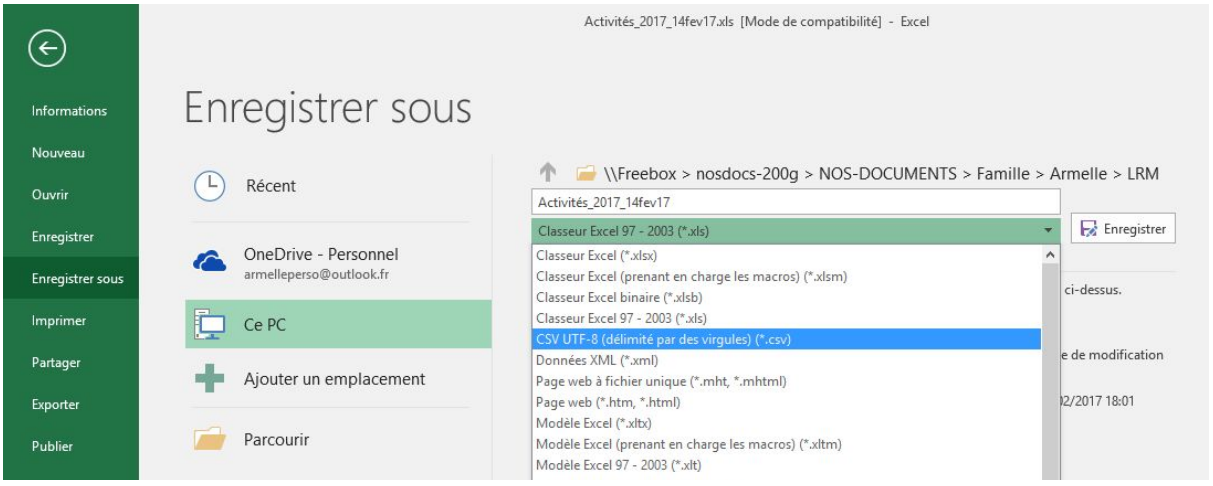
Ce document n'épuise pas ce sujet mais, nous l'espérons, vous aura permis de trouver les bases d'un fichier CSV de qualité.
Pour un public averti, il est possible d'aller plus loin et de mettre en oeuvre une solution très puissante de contrôle qualité et de documentation : le schéma de données. Il s'agit de décrire les données de telle manière qu'un outil de contrôle comme CSV Lint, sera capable de valider automatiquement, pour partie, la qualité des données. Cette technique fera l'objet d'une documentation ultérieure, mais

pour les impatients vous pouvez consulter la norme *Table Schema* : http://specs.frictionlessdata.io/table-schema/

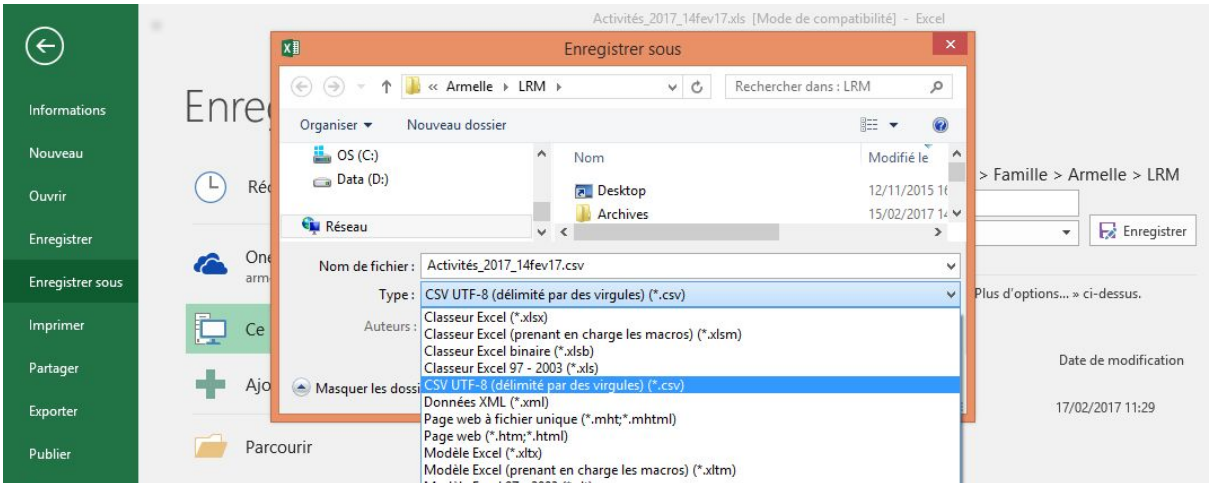
Un CSV de qualité avec Excel

Nous avons utilisé Excel 2016 pour rédiger ce court tutoriel.

- J'ouvre avec Excel mon fichier à convertir
- Je clique sur le menu "**Fichier**" en haut à gauche
- J'ai une interface spécifique qui s'affiche, où je peux sélectionner "**Enregistrer sous…**" (colonne de gauche)
- Dans la partie principale centrale haute de l'écran je peux alors sélectionner, à l'aide d'un menu déroulant, les différents formats de fichier dont "**CSV UTF-8 (délimité par des virgules) (*.csv)**"



- Si je veux changer le fichier d'emplacement, je clique sur le menu "**Parcourir**" grâce auquel j'obtiens une nouvelle fenêtre dans laquelle je peux préciser l'emplacement, mais aussi le type que je sélectionne dans un menu déroulant comme précédemment : "**CSV UTF-8 (délimité par des virgules) (*.csv)**"

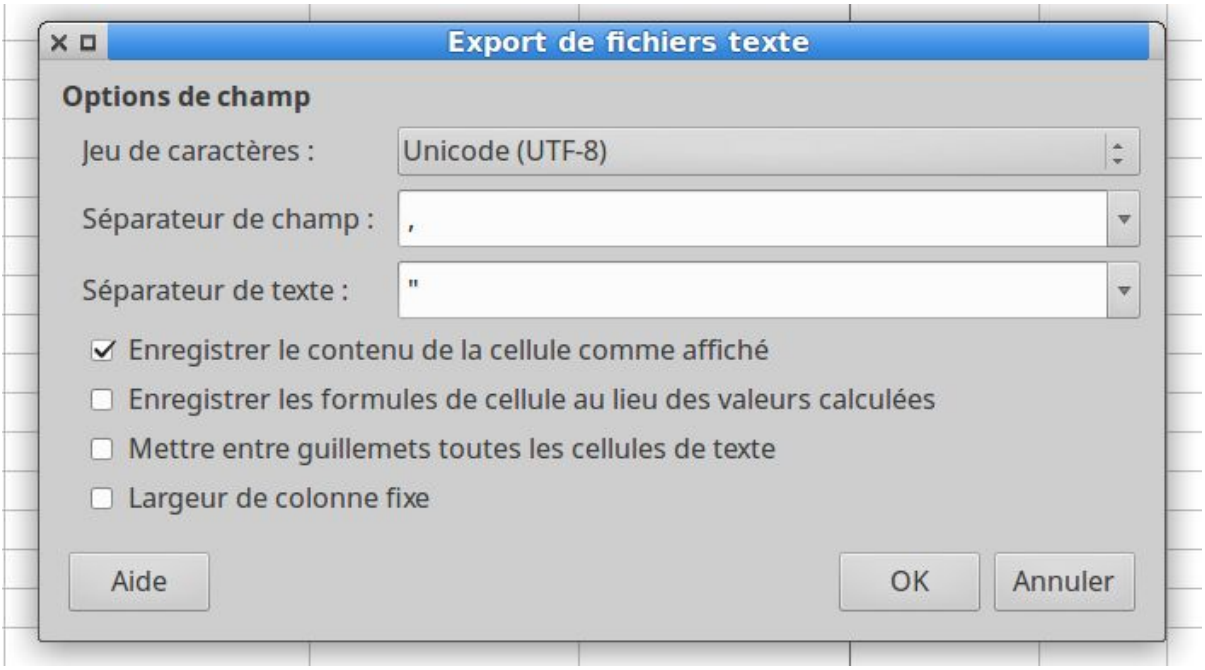


- Une fois enregistré, je n'oublie pas de tester mon fichier avec CSV Lint, afin d'ajuster d'éventuels petits problèmes : http://csvlint.io
- Mon fichier est prêt ! Je peux le publier avec sa documentation.

Un CSV de qualité avec LibreOffice Calc

Nous avons utilisé LibreOffice 5.1 (2016) pour rédiger ce court tutoriel.

- J'ouvre avec LibreOffice Calc (ou OpenOffice Calc) mon fichier à convertir
- Je sélectionne le menu "**Fichier > Enregistrer sous...**"
- Je sélectionne "[x] Éditer les paramètres du filtre" et "Texte CSV (.csv)" dans le menu déroulant des formats.
- S'ouvre alors la boîte de dialogue suivante qui permet de choisir l'encodage et le séparateur : je choisi comme indiqué ci-dessous



- Une fois enregistré, je n'oublie pas de tester mon fichier avec CSV Lint, afin d'ajuster d'éventuels petits problèmes : http://csvlint.io
- Mon fichier est prêt ! Je peux le publier avec sa documentation.