CGIAR Platform for Big Data in Agriculture

CGIAR RESEARCH PROGRAM ON Grain Legumes and Dryland Cereals

CGIAR RESEARCH PROGRAM ON Livestock

General Dataset Curation Guide (GDCG)

Partie I

Curation des jeux de données pour faciliter leur utilisation et leur réutilisation (en utilisant Microsoft Excel)

MEL monitoring evaluation and learning

ICARDA Science for resilient livelihoods in dry areas

GEOAGRO

Etabli en 1977, le Centre International de Recherche Agricole dans les Zones Arides (ICARDA) est un centre de recherche à but non lucratif du CGIAR, qui s'efforce de fournir des solutions innovantes au développement agricole durable dans les zones arides non tropicales des pays en développement.

Nous proposons des solutions novatrices, fondées sur la science, pour améliorer les moyens de subsistance et la résilience des petits exploitants pauvres en ressources. Nous le faisons par le biais de partenariats stratégiques liant la recherche au développement et au développement des capacités, en tenant compte de l'égalité des sexes et du rôle de la jeunesse dans la transformation des zones arides non tropicales.

Pour plus d'informations, s'il vous plaît consultez:
Site principal: http://www.icarda.org/

AUTEURS
Francesco Bonechi[1]

CO-AUTEURS
Enrico Bonaiuti[1], Valerio Graziano[1] et Jane Poole[2]

CITATION SUGGÉRÉE
Francesco Bonechi, Enrico Bonaiuti, Valerio Graziano (2018). General Dataset Curation Guide (GDCG). International Center for Agricultural Research in Dry Areas, Amman, Jordan.

---

[1] International Center for Agricultural Research in the Dry Areas (ICARDA)
[2] International Livestock Research Institute (ILRI)

Historique des revisions

Version	Date	Auteur(s)	Réviseur(s)	Description
1.0	01/10/2018	Francesco Bonechi	Enrico Bonaiuti, Valerio Graziano	Structure, contenu, mise en page
2.0	05/12/2018	Francesco Bonechi	Enrico Bonaiuti, Valerio Graziano	Structure, contenu, mise en page
3.0	15/01/2019	Francesco Bonechi	Enrico Bonaiuti, Valerio Graziano	Structure, contenu, mise en page
4.0	09/06/2019	Francesco Bonechi	Jane Poole	English editing
5.0	18/07/2019	Francesco Bonechi	Valerio Graziano	Structure, branding
6.0	30/10/2019	Francesco Bonechi	Enrico Bonaiuti, Valerio Graziano	Final structural review
7.0	31/11/2019	Francesco Bonechi	Asma Jeitani	Traduction en Français

Remerciements

Lorsque nous parlons de la curation de jeu de données, l'essentiel est toujours de trouver des personnes intéressées à partager leurs jeux de données et de participer à la curation de ces derniers puisqu'il existe certains aspects que seuls les auteurs peuvent connaître avec certitude. Pour cette raison, je voudrais remercier le personnel de l'ICARDA pour sa disponibilité et sa coopération, en fournissant également un retour d'informations utiles pour obtenir de meilleurs résultats. C'est grâce à cet engagement mutuel qu'il a été possible d'achever la curation des différents jeux de données et d'améliorer l'expérience nécessaire à l'élaboration de ce guide.

Je voudrais également remercier sincèrement Enrico Bonaiuti, qui m'a toujours soutenu et supervisé tout au long de ce travail fournissant des suggestions significatives pour progresser dans ce domaine, Valerio Graziano qui, grâce à son expérience, m'aide à structurer ce guide, Jane Poole pour ses précieux commentaires et sa révision, et Zainab Azough pour ses conseils sur des outils automatiques intéressants pour faciliter le formatage des fichiers.

Je suis également reconnaissant au CRP Big Data en Agriculture, au CRP légumineuses à grains et céréales des zones arides, CRP Livestock, MEL et GEOAGRO pour le soutien financier reçu et pour avoir rendu possible ma présence au «Big Data Course» de Rabat et à la «Data Curation» journée à Amman. C'étaient des occasions très précieuses pour améliorer mon expertise et échanger des connaissances avec les autres participants sur ce sujet.

Table des Matières

Acronymes

CRP	CGIAR Research Program
CSV	Comma-separated Value
ICARDA	International Center for Agricultural Research in the Dry Areas
ISO	International Organization for Standardization
MEL	Monitoring, Evaluation and Learning
NA	Not Available
URI	Uniform Resource Identifier
WGS84	World Geodetic System 1984

Introduction

Recruté en tant que consultant pour la plate-forme Big Data, Francesco Bonechi a aidé les scientifiques de l'ICARDA à structurer des jeux de données dans différentes disciplines afin de leur permettre de réutiliser leurs données de recherche au fil du temps. Il les a principalement aidés à développer des fichiers de jeux de données structurés respectant les normes de lisibilités par machine, à créer leur propre dictionnaire de données et à organiser les jeux de données en conservant l'identité et la signification originales de leur contenu. L'objectif était de disposer de bons exemples de jeux de données organisés pour l'ICARDA et de produire un court guide pour les scientifiques, afin d'améliorer la qualité des jeux de données avant de les téléverser sur MEL[3] et par conséquent de générer un lien persistant (Handle[4]) sur le référentiel Dataverse[5].

Processus de la curation des jeux de données

La collecte et l'organisation des données sont l'une des tâches principales des activités de recherche. En fait, la plupart des résultats des projets dépendent de la bonne gestion des données. Cependant, "la valeur à long terme des données peut être affectée, positivement ou négativement, par la qualité de la curation de ces données. Malheureusement, de nombreux jeux de données importants sont mal organisés, ce qui contribue aux erreurs, aux efforts redondants et aux obstacles à la réplication et à l'utilisation" (Ruggles, 2018). En effet, il est courant d'organiser les données dans des feuilles de calcul, de manière facilement compréhensible par l'auteur du jeu de donnée à ce moment-là, sans respecter les normes applicables aux documents lisibles à la machine ni considérer leurs usages possibles dans les futures travaux de recherche. Pour cette raison, il est nécessaire de réviser et d'ajuster ces jeux de données de manière appropriée.

«Les activités de curation de données permettent la découverte et la récupération de données, maintiennent la qualité des données, apportent une valeur ajoutée et permettent leur réutilisation dans le temps» (DH Curation Guide, 2017). Il sera donc important pour quiconque de posséder des connaissances de base sur ce sujet pour pouvoir, lors des activités de recherche, créer de manière autonome des jeux de données bien organisés.

[3]"Monitoring, Evaluation and Learning (MEL) is multi-center and multi-CRP online platform for integrated management, monitoring, and reporting of projects" (GLDC, 2019).
[4] " The Handle System is the Corporation for National Research Initiatives's proprietary registry assigning persistent identifiers, or handles, to information resources, and for resolving those handles into the information necessary to locate, access, and otherwise make use of the resources" (Handle System, 2019). The handle provides the basic framework for the Digital Object Identifier (DOI) system that became the official ISO standard in 2012 (ISO 26324).
[5] "Dataverse is an open source web application to share, preserve, cite, explore, and analyze research data" (The Dataverse Project, 2019).

Le guide élaboré aborde certaines étapes de la curation des jeux de données: nettoyage de fichiers, créer un dictionnaire de données faisant référence au vocabulaire standard et à convertir les fichiers à partir d'un format de logiciel sous licence (Microsoft Excel, par exemple), qui ne peut être ouvert qu'en utilisant la même famille de produits avec des versions prises en charge spécifiques, vers un format stable comme le format CSV, compatible avec de nombreux produits sous licence et source-ouverte, logiciel d'analyse statistique inclus, garantissant que le fichier peut être lu à l'avenir. De cette manière, le jeu de données durera bien au-delà de la portée actuelle, tout en conservant sa validité aux fins de la recherche.

Références utiles pour des lectures supplémentaires:

- "Data Carpentry: Data Organization in Spreadsheets Ecology lesson" (Bahlai, 2017);structurer les fichiers en fonction des normes de lisibilités par machine;
- "Ag Data Commons Data Submission Manual v1.3" (USDA, 2016); développer et gérer le dictionnaire de données des jeux de données.

Outil utile pour prendre en charge le nettoyage et la conversion des fichiers pour établir la lisibilité par machine:

- «Talend Data Preparation» est l'un des logiciels testés capable d'identifier automatiquement les erreurs dans les jeux de données et de permettre des actions de nettoyage et de formatage (lien disponible dans l'Annexe B).

1. Étapes Rréliminaires

«Lorsque vous travaillez avec des feuilles de calcul, lors du nettoyage ou l'analyse des données, il est très facile de créer une feuille de calcul très différente de celle avec laquelle vous avez commencé» (Bahlai, 2017). Afin de pouvoir reproduire vos analyses et d'être sûr de ne perdre aucune donnée dans les processus de curation, ne modifiez pas le jeu de données original mais créez une nouvelle copie dans laquelle vous organisez votre arrangement de données.

2. Calculs et Résumés de Données

«Les feuilles de calculs sont utiles pour la saisie de données, mais il est courant d'utiliser des tableurs pour bien plus que cela. Ils servent à créer des tableaux de données pour les publications, à générer des statistiques résumées et à faire des chiffres » (Bahlai, 2017). Mais tout cela ne doit pas être inclus dans le jeu de données organisé stsndard.

En fait, le jeu de données ne devrait contenir que les données brutes nettoyées. Toute autre élaboration (tableaux croisés dynamiques, graphiques, formules, etc.) ne doit pas être présente car elle représente les étapes de recherche successives et ne doit pas affecter l'intégrité du jeu de données réel (Fig. 2.1 et 2.2). Les développements supplémentaires seront ensuite présentés avec les prochaines publications générées à partir de l'analyse et de l'étude du jeu de données.

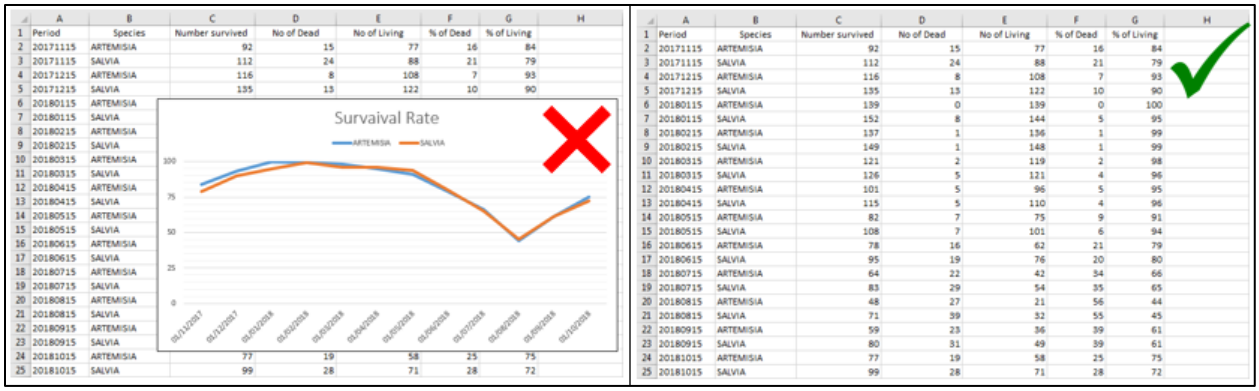


Figure 2.1. *Les graphiques et les figures doivent être supprimés de l'onglet de la feuille de calcul du jeu de données.*

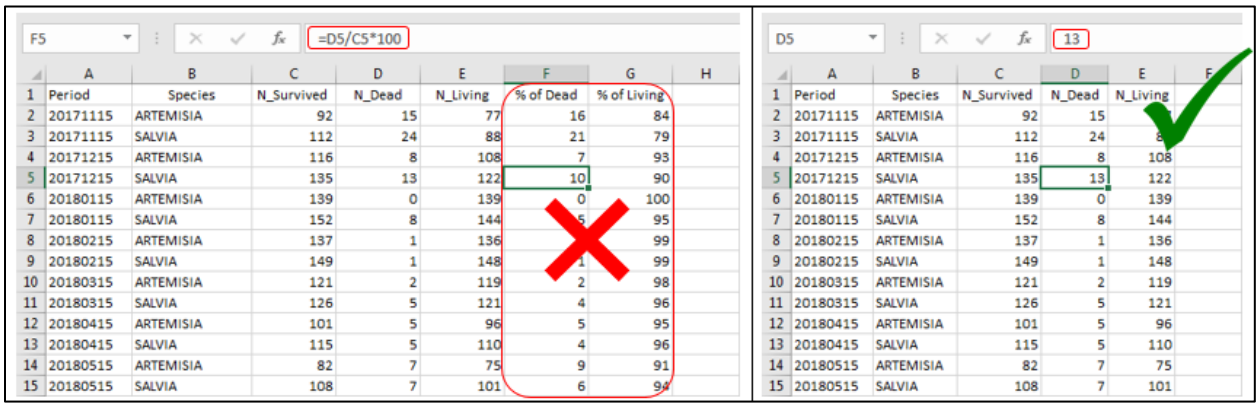


Figure 2.2. *Les formules et tout autre type d'élaboration doivent être supprimés de l'onglet de la feuille de calcul. Cela peut forcer la suppression de colonnes entières de la feuille de calcul.*

Remarque: dans certains cas, les données sont collectées pour effectuer des calculs spécifiques (par exemple, poids net,% de la production, etc.). Cependant, même si les résultats des calculs constituent l'essentiel des activités de recherche, ceux-ci ne doivent pas affecter le jeu de données d'origine. Leur inclusion peut influencer l'utilisation des données par d'autres.

3. Tableau de données

Une autre étape importante consiste à ajuster les tableaux de la feuille de calcul. Il n'est pas possible d'avoir plusieurs tables de données dans une même feuille de calcul et d'utiliser des lignes ou des colonnes vierges pour séparer les données. Ceci parce que l'ordinateur n'est pas capable de distinguer les différentes tables créant des associations fausses (Bahlai, 2017).

Comme le montre la figure 3.1, les tableaux supplémentaires doivent être coupés de la feuille active et collés dans un nouvel onglet. À la fin de cette opération, vous aurez plus d'onglets dans le fichier. Chaque onglet contiendra l'un des tableaux précédemment disponibles dans la même feuille. Le résultat est une table pour chaque onglet.

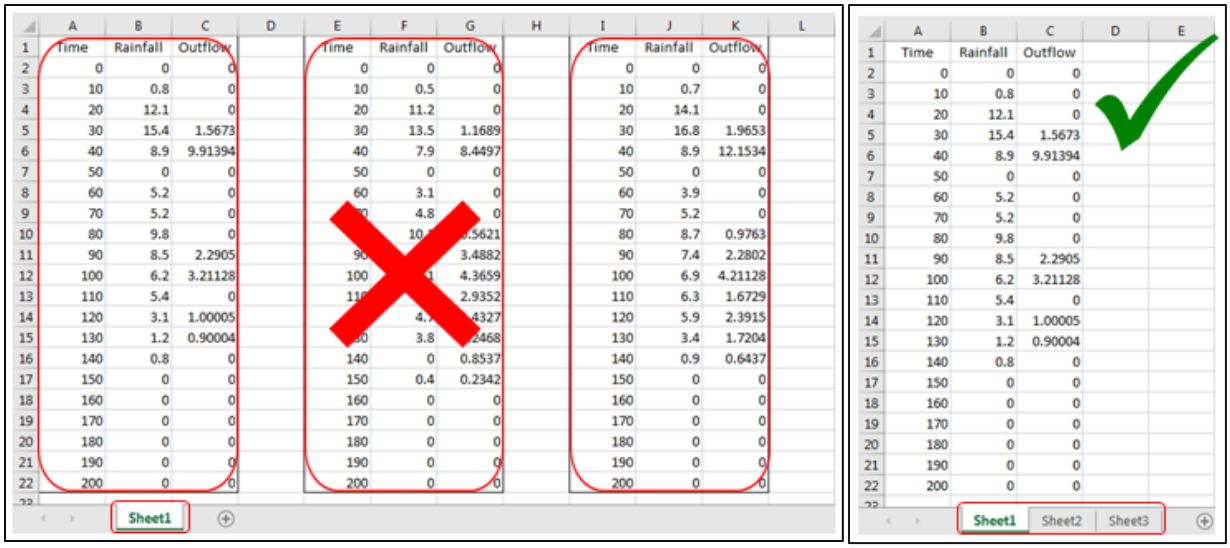


Figure 3.1. *Afin d'éviter une association erronée lors de la lisibilité du système de données, vous ne devez pas utiliser de lignes ni de colonnes vierges pour séparer le jeu de données dans différentes tableaux ou sections.*

Remarque: Dans cet exemple, nous créons plusieurs onglets pour organiser le fichier de données. Cependant, «lorsque vous créez des onglets supplémentaires, vous ne permettez pas à l'ordinateur de voir les connexions dans les données et vous êtes plus susceptible d'ajouter accidentellement des incohérences dans le fichier" (Bahlai, 2017). Pour cette raison, s'il existe un lien entre les différentes feuilles, vous devez les combiner en un seul. Cela peut arriver, par exemple, lorsque vous créez un onglet distinct pour chaque jour où vous prenez une mesure. Ce problème peut être résolu en ajoutant la colonne «Date», en évitant toute répétition d'onglets (Fig. 3.2).

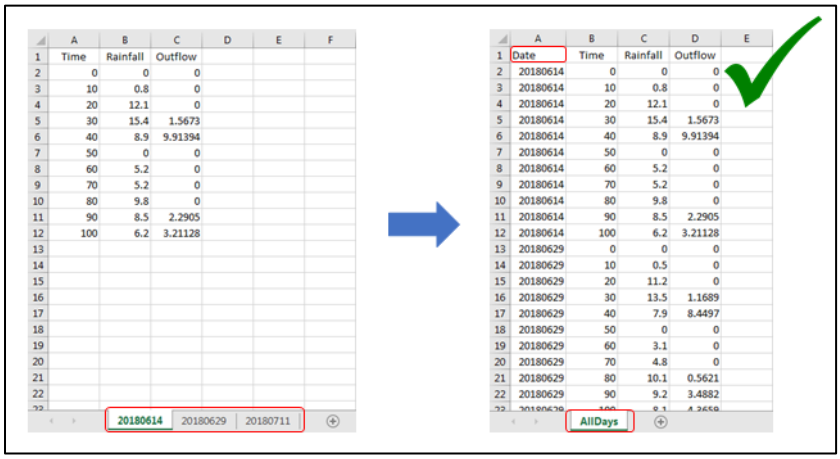


Figure 3.2. *Il est possible d'améliorer la disposition des colonnes du jeu de données afin de réduire le nombre d'onglets de la feuille de calcul.*

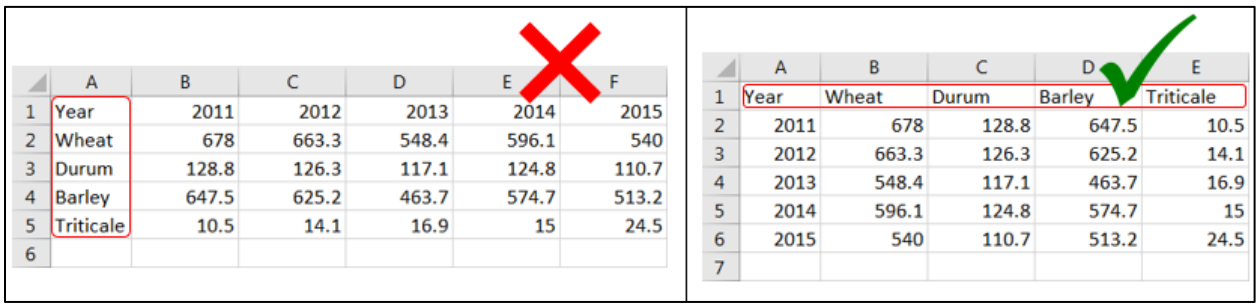
4. Structuration des Données dans une Feuille de Calcul

Une fois la manipulation des données basique est effectuée, nous pouvons procéder à la structuration correcte des données de la feuille de calcul.

«Les règles fondamentales d'utilisation des tableurs pour les données sont» (Bahlai, 2017):

- Placez toutes les variables en colonnes. Chaque colonne correspond à une variable.
- Placez chaque observation dans sa propre ligne. Chaque ligne correspond à une observation.
- Ne combinez pas plusieurs informations dans une cellule. Chaque cellule correspond à une valeur (donnée).

Afin de respecter ces principes, les données doivent être organisées selon un schéma fixe, les noms des champs (ou des variables) correspondant aux en-têtes de colonnes et les différentes observations étant organisées dans les lignes correspondantes (Fig. 4.1).



	A	B	C	D	E	F
1	Year	2011	2012	2013	2014	2015
2	Wheat	678	663.3	548.4	596.1	540
3	Durum	128.8	126.3	117.1	124.8	110.7
4	Barley	647.5	625.2	463.7	574.7	513.2
5	Triticale	10.5	14.1	16.9	15	24.5
6						

	A	B	C	D	E
1	Year	Wheat	Durum	Barley	Triticale
2	2011	678	128.8	647.5	10.5
3	2012	663.3	126.3	625.2	14.1
4	2013	548.4	117.1	463.7	16.9
5	2014	596.1	124.8	574.7	15
6	2015	540	110.7	513.2	24.5
7					

Figure 4.1. *À gauche, une organisation des données non correcte. À droite, l'agencement de données suggéré, où les colonnes correspondent aux variables, les lignes aux observations et les cellules aux valeurs.*

Une fois la structure du jeu de données est passée en revue, des actions supplémentaires pour une curation complète sont indiquées ci-dessous.

4.1. Fonctions de Mise en Forme

Les mises en forme et fonctionnalités spéciales (cellules fusionnées, frontières, couleurs, en gras, etc.) doivent être évitées autant que possible. En fait, tous ces aspects facilitent l'approche humaine des données mais créent de nombreux problèmes pour les processus lisibles par machine. Chaque jeu de données doit avoir une structure simple de colonnes et de lignes. «Pensez à restructurer vos données de manière à ne pas devoir fusionner des cellules ou utiliser autres éléments esthétiques pour organiser vos données» (Bahlai, 2018) (Fig. 4.2).

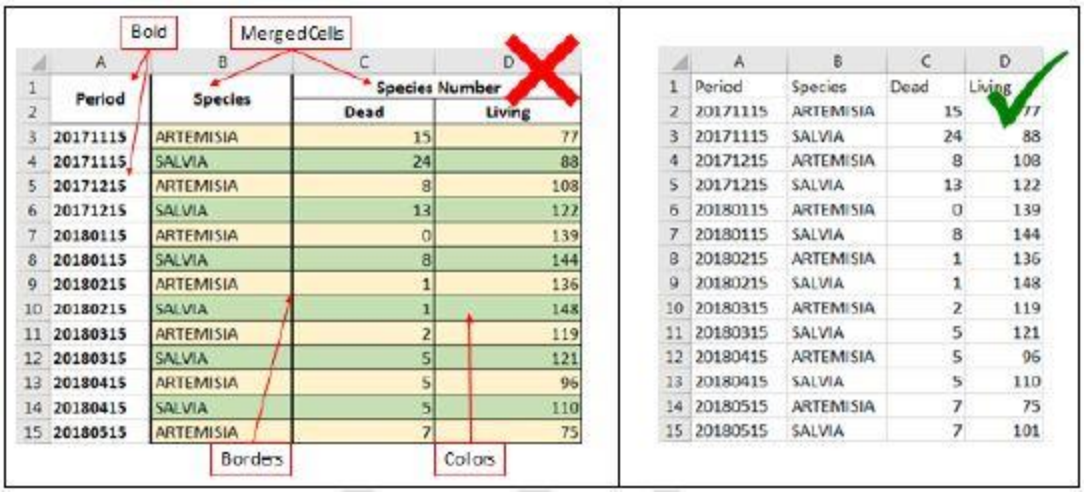


Figure 4.2. *Les mises en forme et fonctionnalités spéciales sont supprimées de l'onglet pour faciliter les prochains processus de lisibilité par machine.*

4.2. En-têtes de colonnes

N'ajoutez pas la documentation et les descriptions textuelles dans les tableaux eux-mêmes. Les informations descriptives peuvent être enregistrées dans le dictionnaire de données ou insérées dans une colonne «Note» créée à cet effet (USDA, 2017). Les en-têtes de colonnes doivent indiquer le contenu de chaque colonne sans description supplémentaire. «Considérez une longueur limitée de vos noms de variables. La plupart des logiciels lisent des noms courts. Pour cette raison, il est suggéré d'utiliser des noms de variables ne dépassant pas les 8 caractères, commençant par une lettre » (IITA, 2019). Ensuite, assurez-vous que les en-têtes de colonnes ne contiennent pas d'espaces, de traits d'union ou d'autres symboles. Seul le trait de soulignement est autorisé (Fig. 4.3).

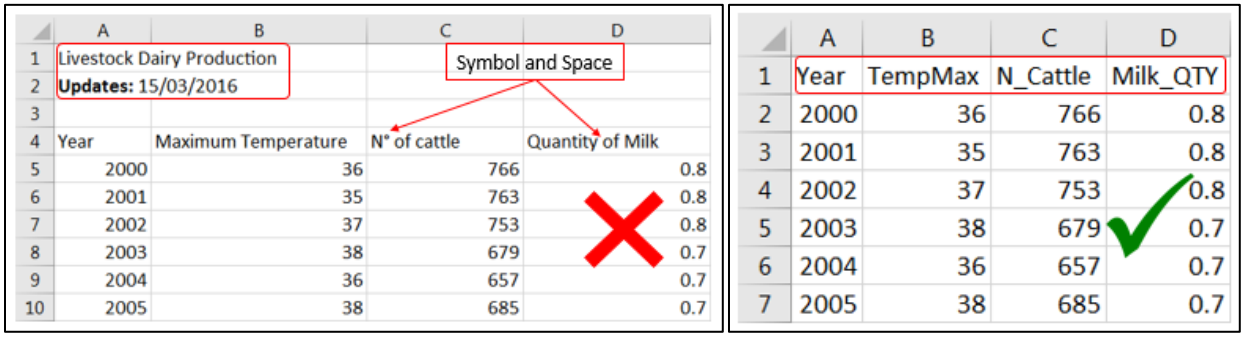
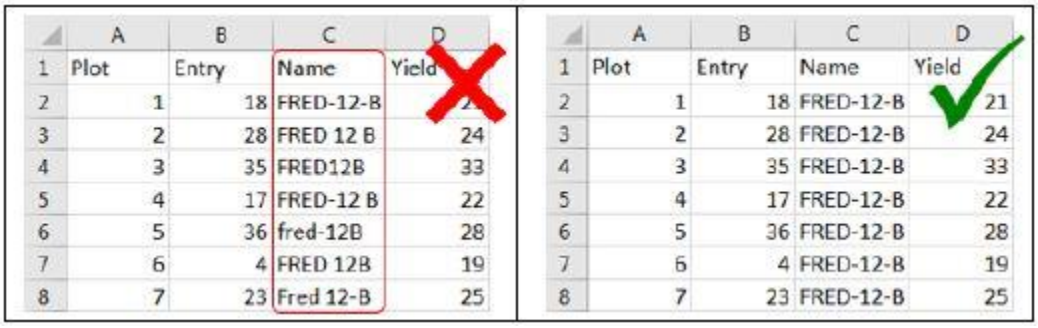


Figure 4.3. *Les documentations supplémentaires sont supprimées de l'onglet et l'en-tête de colonne est écrit de manière cohérente.*

Remarque: " Les tirets bas (_) sont une bonne alternative aux espaces. Envisagez d'écrire les noms en camelcase (par exemple, Nom du test) pour améliorer la lisibilité » (Bahlai, 2017).

4.3. Saisie de données

Les données doivent être saisies de manière cohérente en utilisant toujours le même code pour la même valeur. En fait, les codes doivent être écrits avec soin, en prenant soin de les écrire en utilisant toujours le même format en termes d'espaces, de symboles et d'autres caractéristiques adoptées (Fig. 4.4).



	A	B	C	D
1	Plot	Entry	Name	Yield
2	1	18	FRED-12-B	21
3	2	28	FRED 12 B	24
4	3	35	FRED12B	33
5	4	17	FRED-12 B	22
6	5	36	fred-12B	28
7	6	4	FRED 12B	19
8	7	23	Fred 12-B	25

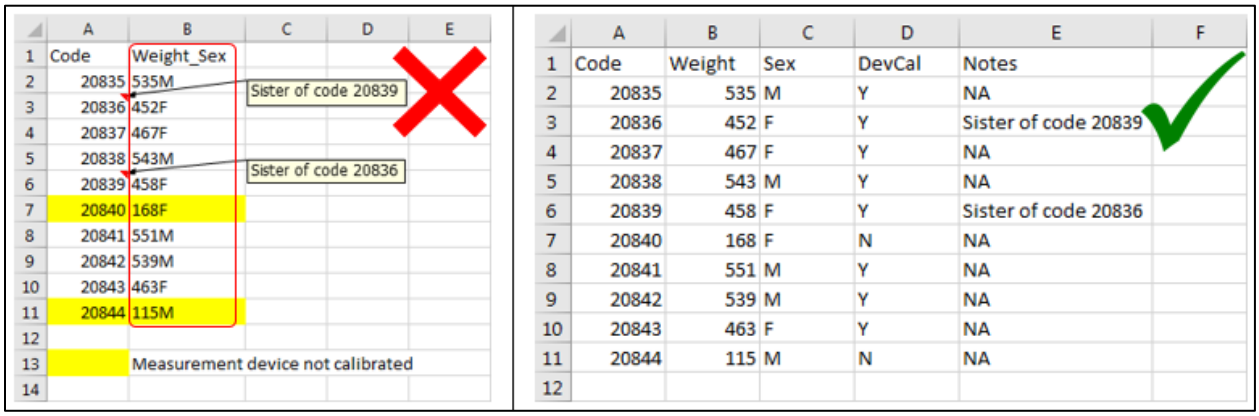
	A	B	C	D
1	Plot	Entry	Name	Yield
2	1	18	FRED-12-B	21
3	2	28	FRED-12-B	24
4	3	35	FRED-12-B	33
5	4	17	FRED-12-B	22
6	5	36	FRED-12-B	28
7	6	4	FRED-12-B	19
8	7	23	FRED-12-B	25

Figure 4.4. *Le même code est saisi sans utiliser un format cohérent (espaces et symboles) à gauche et le même code est saisi de manière cohérente à droite.*

Pas plus d'une information ne peut être dans une cellule. Si d'autres détails de mesure doivent être ajoutés, ils doivent être saisis dans des colonnes supplémentaires en laissant les valeurs séparées. Ceci afin d'éviter des problèmes lors de l'analyse suivante du jeu de données et de garder l'ensemble de la structure propre.

Ceci s'applique également aux cellules et commentaires mis en surbrillance. En fait, même s'il est courant d'utiliser ces fonctionnalités pour ajouter des notes aux données, celles-ci doivent être supprimées car elles peuvent créer des problèmes de lisibilité par machine. Ces observations peuvent être entrées dans de nouvelles colonnes. Il est également possible de créer une colonne «Notes» où consigner les commentaires ou d'autres informations (Fig. 4.5).

Il n'est pas nécessaire d'inclure des unités dans les cellules. Celles-ci seront rapportées dans le dictionnaire de données (description des éléments du jeu de données). Toutefois, si plusieurs unités différentes peuvent être utilisées lors de la collecte de données et que vous devez saisir ces informations dans le jeu de données, envisagez d'ajouter une colonne supplémentaire pour enregistrer ces données.



	A	B	C	D	E
1	Code	Weight_Sex			
2	20835	535M			
3	20836	452F			
4	20837	467F			
5	20838	543M			
6	20839	458F			
7	20840	168F			
8	20841	551M			
9	20842	539M			
10	20843	463F			
11	20844	115M			
12					
13		Measurement device not calibrated			
14					

Sister of code 20839

Sister of code 20836

	A	B	C	D	E	F
1	Code	Weight	Sex	DevCal	Notes	
2	20835	535	M	Y	NA	
3	20836	452	F	Y	Sister of code 20839	
4	20837	467	F	Y	NA	
5	20838	543	M	Y	NA	
6	20839	458	F	Y	Sister of code 20836	
7	20840	168	F	N	NA	
8	20841	551	M	Y	NA	
9	20842	539	M	Y	NA	
10	20843	463	F	Y	NA	
11	20844	115	M	N	NA	
12						

Figure 4.5. *À gauche, un ensemble de données avec des commentaires, les cellules mises en surbrillance et la colonne «B» contenant plusieurs informations dans une cellule (poids et sexe). Sur la droite, le jeu de données organisé avec les colonnes supplémentaires pour saisir toutes les différentes valeurs afin de faciliter l'analyse du jeu de données.*

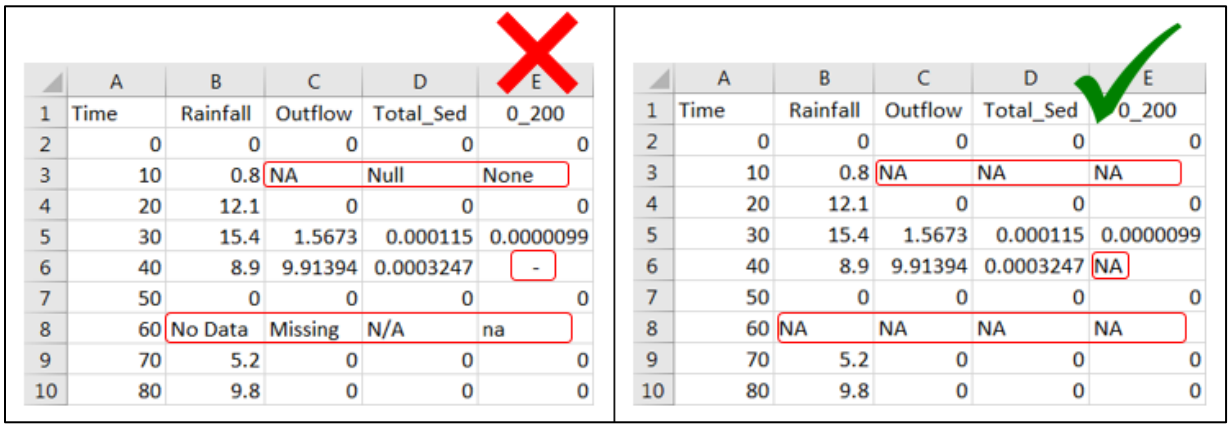
Remarque: lors de la rédaction de texte dans les cellules (comme pour la colonne «Notes» de la Fig. 4.5), elles ne peuvent contenir que du texte et des espaces. Cela signifie que l'ajout de caractères tels que ceux de fin de ligne, les tabulations et les tabulations verticales doit être évité (Bahlai, 2017).

4.4. Valeurs Nulles

Les valeurs NULL doivent être représentées différemment de «0». En fait, «0» correspond à une donnée mesurée alors que la valeur nulle signifie que la donnée n'a pas du tout été mesurée. En ne saisissant pas la valeur de votre observation, l'ordinateur interprétera ces données comme inconnues ou manquantes (null). "Pour cette raison, il est très important d'enregistrer les zéros comme zéros et les données réellement manquantes comme nulles" (Bahlai, 2017).

Cependant, les valeurs nulles ou les informations manquantes peuvent être représentées différemment dans les différents jeux de données. En effet, chaque programme statistique nécessite un style spécifique pour les accepter et les lire. Par conséquent, la manière dont ces valeurs doivent être consignées dans le jeu de données dépend du logiciel utilisé pour analyser les données.

Il est courant de représenter des valeurs nulles avec des cellules vides, alternativement «NA» ou «NULL» sont de bonnes options (White, 2013). L'utilisation de valeurs numériques (par exemple 999, -999) ou d'autres codes et indications textuelles (par exemple, Missing, No data, None, -, +) est une autre possibilité pour indiquer des valeurs nulles ou manquantes. Ces valeurs ne sont toutefois pas recommandées puisqu'ils peuvent causer des problèmes liés à l'utilisation de plusieurs logiciels (White, 2013). Dans tous les cas, il est essentiel de sélectionner un indicateur nul clair et cohérent qui doit être défini dans les descriptions de métadonnées (par exemple, le dictionnaire de données) (Fig.4.6).



	A	B	C	D	E
1	Time	Rainfall	Outflow	Total_Sed	0_200
2	0	0	0	0	0
3	10	0.8	NA	Null	None
4	20	12.1	0	0	0
5	30	15.4	1.5673	0.000115	0.0000099
6	40	8.9	9.91394	0.0003247	-
7	50	0	0	0	0
8	60	No Data	Missing	N/A	na
9	70	5.2	0	0	0
10	80	9.8	0	0	0

	A	B	C	D	E
1	Time	Rainfall	Outflow	Total_Sed	0_200
2	0	0	0	0	0
3	10	0.8	NA	NA	NA
4	20	12.1	0	0	0
5	30	15.4	1.5673	0.000115	0.0000099
6	40	8.9	9.91394	0.0003247	NA
7	50	0	0	0	0
8	60	NA	NA	NA	NA
9	70	5.2	0	0	0
10	80	9.8	0	0	0

Figure 4.6. *Valeurs manquantes incohérentes à gauche et valeurs nulles correctement consignées à droite.*

Remarque: Les représentations explicites des valeurs manquantes sont préférables aux champs vides, mais cette exigence dépend du logiciel utilisé pour analyser les données. En fait, en particulier pour les champs de date, le travail avec des valeurs nulles a ses propres complications car la plupart des bases de données n'autorisent pas la valeur nulle pour une date. Cela signifie que si la base de données que nous allons utiliser pour analyser les données est déjà connue, la décision de représenter les données manquantes lors de la curation du jeu de données doit être prise basée sur la valeur acceptée de cette base de données pour représenter la valeur nulle dans une variable de date. Cependant, tant que la représentation de la valeur manquante est cohérente et documentée, les utilisateurs suivants peuvent remplacer le choix d'une valeur null de manière indépendante (Zwicker, 2016).

4.5. Date et heure

Il est courant que les données saisies sous forme de dates et d'heures dans la feuille de calcul soient gérées par des fonctions spécifiques aux logiciels, qui représentent ces données conformément à certaines normes par défaut du programme. Cela peut créer des ambiguïtés dans les jeux de données et un problème pour les utilisations suivantes. Pour éviter ces problèmes, les fonctionnalités spéciales disponibles ne doivent pas être utilisées car elles ne sont généralement compatibles qu'avec la même famille de produits (Fig. 4.7) (Bahlai, 2017).

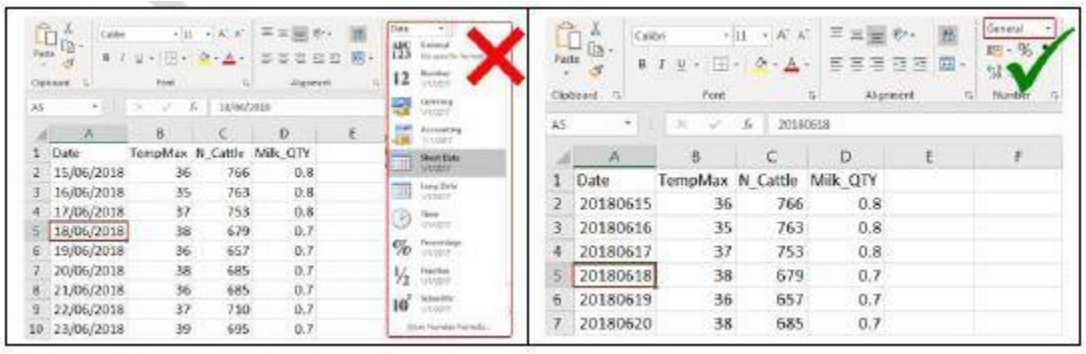


Figure 4.7. *Adoption du format spécifique au logiciel à gauche et adoption d'aucun format spécifique au logiciel à droite.*

Pour cette raison, basé sur les normes ISO (ISO 8601: 2004), le format suggéré pour stocker les dates est AAAAMMJJ, tandis que pour l'heure c'est hhmmss utilisant la notation sur 24 heures (qui devient AAAAMMJJhhmmss lorsqu'elles sont représentées ensemble). «Par exemple, le 24 mars 2015 à 17 h 25 devenant 20150324172535. Ces chaînes seront correctement triées par ordre croissant ou décroissant et, en connaissant le format, elles peuvent ensuite être traitées correctement par le logiciel récepteur» (Bahlai, 2017).

Une autre option pour supprimer toute ambiguïté dans le jeu de données consiste à stocker les valeurs d'années, mois, jours, heures, minutes et secondes dans différentes colonnes. En fait, «traiter les dates comme plusieurs données plutôt qu'une seule les rend plus faciles à manipuler» (Bahlai, 2017).

4.6. Latitudes et Longitudes

Un aspect important de la collecte de données consiste à suivre la position de l'endroit où les mesures ont été prises. Cependant, dans la représentation des informations de position globale, il existe plusieurs façons de rapporter des données de latitude et de longitude. Parmi ceux-ci, la norme recommandée pour leur représentation est l'utilisation de degrés décimaux (DD), car ils garantissent la possibilité de traiter la latitude et la longitude comme une valeur simple et numérique facilitant les interprétations logicielles suivantes (Callahan, 2009).

Ainsi, basé sur la norme proposée (Callahan, 2009):

- Les latitudes sont stockées sous forme de valeurs numériques dans la plage de [-90,90] avec des unités de degrés décimaux. Les valeurs positives indiquent l'hémisphère Nord alors que les valeurs négatives l'hémisphère Sud.
- Les longitudes sont stockées sous forme de valeurs numériques dans la plage [-180,180] avec des unités de degrés décimaux. Les valeurs positives indiquent l'hémisphère oriental tandis que les valeurs négatives l'hémisphère occidental.

Unit	Latitude	Longitude
Decimal Degrees (DD)	**40.75889**	**-73.98513**
Degrees Minutes and Seconds (DMS)	40° 45' 32.004" N	73° 59' 6.468" W
Degrees Decimal Minutes (DM)	40° 45.5334'	-73° 59.1078'

	A	B	C
1	Site	Latitude	Longitude
2	PT-12A	39.91382	116.36363
3	PT-34B	40.75889	-73.98513
4	PT-41C	-22.90278	-43.20750
5	PT-56D	-33.86785	151.20732

Figure 4.8. *À gauche, en gras, la norme suggérée pour la représentation des coordonnées. A droite, latitude et longitude exprimées en degrés décimaux dans un ensemble de données.*

5. Formats de fishiers stables

Une fois toutes les procédures mentionnées précédemment sont achevées, le jeu de données doit maintenant être enregistré dans un format de fichier stable. En fait, lors de l'enregistrement des fichiers sous un format de logiciel sous licence tel que Microsoft Excel, il est possible que les documents ne s'ouvrent pas avec un autre logiciel ou même avec Excel lui-même si le jeu de données a été créé en adoptant une version plus ancienne qui n'est plus prise en charge. Pour cette raison, il est important d'enregistrer le jeu de données dans un format cohérent, lisible dans le futur et indépendant des modifications apportées aux applications.

Les fichiers de valeurs séparées par des virgules ou CSV constituent le format de données préféré pour la plupart des référentiels de données et sont recommandés pour la publication de données tabulaires lisibles par machine. Cependant, avant de procéder à la conversion, il est important de vérifier les onglets du fichier Excel. En effet, il n'est pas possible d'enregistrer au format CSV une feuille de calcul contenant plusieurs onglets. Ainsi, si le fichier Excel du jeu de données contient plusieurs onglets, ceux-ci doivent être séparés en différents fichiers. Par exemple, si le fichier Excel du jeu de données contient 5 onglets différents, à la fin de cette opération, vous aurez 5 fichiers Excel différents contenant chacun un onglet. En résumé, une feuille de calcul multi-onglets deviendra plusieurs fichiers.

Désormais, les fichiers Excel avec un seul onglet sont prêts à être convertis au format texte CSV.

5.1. Fichier CSV

CSV est un fichier texte délimité qui utilise une virgule pour séparer les valeurs. Il s'agit d'un format commun d'échange de données largement pris en charge par les applications grand public, commerciales et scientifiques (Comma-separated values, 2018). Cette large applicabilité «signifie que les données au format CSV ont moins de chances de devenir obsolètes en raison de leur inaccessibilité et de leur longévité plus longue que les formats de fichiers sous licence. De plus, les fichiers CSV sont plus polyvalents et lisibles par machine (l'ordinateur peut extraire, transformer et traiter les données) » (USDA, 2016).

En général, la facilité de conversion du fichier Excel au format CSV dépend de l'état actuel des données. L'objectif final est d'avoir une seule page de feuille de calcul avec une ligne unique des en-têtes de colonnes en haut de la page » (USDA, 2016). Toutefois, si les indications données dans les sections précédentes de ce guide ont été suivies, la conversion devrait être un processus rapide et facile. Dans tous les cas, voici quelques points clés qui doivent être vérifiés avant de passer par le processus de conversion (USDA, 2016):

- Les données ont une seule ligne des en-têtes de colonnes pour étiqueter les variables du jeu de données.
- Évitez autant que possible les virgules de votre document. Étant donné que le délimiteur CSV est une virgule, des virgules supplémentaires dans le texte peuvent entraîner des erreurs d'interprétation des données.
- Les données contenant plusieurs tables sont combinées dans une seule table ou séparées en différents onglets (feuilles de calcul) et par conséquent divers fichiers.
- Chaque fichier est une feuille de calcul autonome avec un seul onglet et aucune autre information supplémentaire (même les onglets vierges ont été supprimés du fichier).

Une fois que le tableur a respecté ces normes, enregistrez le fichier au format CSV.
Maintenant, le document enregistré peut être ouvert avec Excel ou tout autre logiciel en fonction des objectifs de l'application de données.

Remarque: à partir de maintenant, le jeu de données est représenté par le dossier contenant tous les fichiers CSV. Pour cette raison, améliorez le nom du dossier «en saisissant un titre descriptif comprenant des dates, des lieux et des mesures spécifiques qui rendent le jeu de données unique» (USDA, 2016). De la même manière, assurez-vous que tous les fichiers à l'intérieur sont nommés avec un titre cohérent et descriptif afin de pouvoir identifier facilement le contenu de leurs données (Hodge, 2015).

6. Dictionnaires de données

«Les dictionnaires de données sont utilisés pour fournir des informations détaillées sur le contenu d'un jeu de données ou d'une base de données, tels que les noms des variables mesurées, les types ou formats des données et les descriptions. Un dictionnaire de données fournit un guide concis pour comprendre et utiliser les données » (USDA, 2016). En outre, la possibilité d'enregistrer ces informations dans un fichier séparé facilite le maintien du jeu de données brutes propre et facile à analyser.

Le dictionnaire de données joue un rôle crucial dans les processus de réutilisation des données et dans la compréhension du contenu du jeu de données. En particulier (Briny, 2015):

- Permet à l'auteur du jeu de données de se rappeler de tous les détails des données au fil des années;
- Facilite le partage des jeux de données avec les collaborateurs en les aidant à comprendre et à utiliser les fichiers de données.
- Utile pour le personnel «qui connaît totalement les données, de récupérer ces données, de comprendre et de reproduire les résultats ou de les réutiliser pour de nouvelles recherches» (Briny, 2015) en améliorant la qualité du jeu de données.

Par conséquent, un dictionnaire de données fait la différence entre avoir un jeu de données réutilisable à des fins de recherche ou non. «Il ne s'agit pas nécessairement d'une documentation sur les données elles-mêmes, mais d'une documentation permettant de donner le contexte dans lequel ces données sont comprises» (Briny, 2015).

En général, lorsque les données sont gérées dans des bases de données professionnelles, il est possible de générer automatiquement des dictionnaires de données à l'aide des outils disponibles dans le logiciel. «Cela fournira un document qui est systématiquement formaté et contient ce dont les autres ont besoin pour comprendre vos données» (USDA, 2016).

Alors que, lorsque les données sont gérées dans des feuilles de calcul, des fichiers texte ou des valeurs séparées par des virgules, le dictionnaire de données doit être créé manuellement. Pour faciliter la lisibilité, il est recommandé de préparer le dictionnaire de données sous forme de feuille de calcul. S'il est préférable de le préparer au format .doc ou .pdf, le tableau dans le document devrait être facilement extractible (USDA, 2016).

Une approche courante lorsque vous le faites manuellement consiste à créer trois fichiers principaux, à enregistrer au format CSV, qui contient trois niveaux d'informations de jeu de données différents:

1. **Dataset Introduction:** une introduction et des explications générales sont fournies;
2. **Dataset Elements Descriptions:** les champs des jeux de données sont répertoriés avec leurs informations associées.
3. **Unique Identifier:** les éléments et les concepts du jeu de données sont identifiés par des URI déréférencables vers un thésaurus multilingue en ligne.

Combinés, ces éléments peuvent être appelés les ‘métadonnées’ du jeu de données. Lorsque vous utilisez Microsoft Excel pour créer les fichiers de dictionnaire de données, vous devez suivre les principes expliqués précédemment pour la structure du jeu de données afin de les enregistrer au format CSV lors de la création de ces documents.

6.1. Introduction au jeu de données

Le but de l'onglet d'introduction du jeu de données est d'expliquer le contenu du jeu de données. Voici les informations générales disponibles pour clarifier tous les aspects du jeu de données pour les utilisations suivantes. Les champs de cet onglet sont (Fig. 6.1):

- **Description:** description libre de texte riche fournissant le plus d'explications possible sur le jeu de données: comment et pourquoi il a été généré et comment il devrait (ou ne devrait pas) être utilisé. Assurez-vous que, dans cette description, sont présents les paramètres de l'expérience (lieux, conditions climatiques, etc.), les méthodes de collecte de données et de processus, l'équipement utilisé, la période, les ressources possibles et les facteurs limitants (USDA, 2016). Il devrait également inclure les éléments de conception qui sont importants pour interpréter les données (par exemple, population cible, stratification, échantillon, taille).
- **Summary:** description plus courte du jeu de données, généralement pas plus d'une phrase ou deux (USDA, 2016).
- **Start_Date:** date à laquelle la collecte de données commence.
- **End_Date:** la date à laquelle la collecte de données se termine.
- **Latitude:** coordonnées du site de latitude à l'aide du système de référence WGS84.
- **Longitude**: coordonnées du site de longitude à l'aide du système de référence WGS84.
- **Author:** premier auteur du jeu de données.
- **CoAuthor:** co-auteurs du jeu de données.

	A	B	C	D	E	F	G	H
1	Description	Summury	Start_Date	End_Date	Latitude	Longitude	Author	CoAuthor
2	A rich free text description that provides as much explanation as possible about the dataset: how and why it was generated, and how it should (or should not) be used. Make sure that in this description are present the experiment settings (location, climatic conditions, etc.), data collection and processes methods, equipment used, period, possible resources and any limiting factors (USDA, 2016).	A shorter description of the dataset, usually no more than a sentence or two (USDA, 2016).	The date in which the data collection starts (YYYYMMDD).	The date in which the data collection ends (YYYYMMDD).	Latitude site coordinates using WGS84 reference system.	Longitude site coordinates using WGS84 reference system.	Dataset first author.	Dataset Co-Author.

Figure 6.1. *Champs «DataDictionary_DatasetIntroduction». Une colonne «Notes» peut être ajoutée pour indiquer les caractéristiques pertinentes du jeu de données, comme le code adopté pour exprimer des valeurs nulles (ou des informations manquantes).*

Les champs ci-dessus sont les champs de base et suggérés pour l'onglet «Introduction au jeu de données». Toutefois, s'il reste d'autres informations à déclarer (Author ORCID identifier, etc.), des colonnes supplémentaires peuvent être ajoutées pour enrichir et compléter le formulaire. Il s'agit d'un modèle polyvalent pouvant être adapté aux différents besoins et thèmes du jeu de données.

Veuillez noter que le fichier «Introduction au jeu de données» permet de rapporter les coordonnées d'un seul emplacement. Ainsi, lorsque les données ont été collectées à plusieurs endroits, les détails du site (au moins la latitude et la longitude) doivent être disponibles dans le jeu de données lui-même. Dans le cas contraire, la meilleure option serait de créer un fichier «Résumé du site» où consigner les détails des différentes zones (annexe A, figure A.1).

6.2. Description des éléments du jeu de données

Une fois terminé avec l'introduction du jeu de données, c'est le moment d'expliquer en détail les éléments du jeu de données.

C'est le noyau du dictionnaire de données, car c'est le document qui permet aux utilisateurs du jeu de données de comprendre pleinement son contenu, y compris les noms de paramètre, les unités de mesure, les formats et les définitions de valeurs codées (DAAC ORNL, 2018).

Le modèle suggéré pour structurer manuellement la «Description des éléments de jeu de données» comprend les champs suivants (USDA, 2016):

- **Spreadsheet_Tab:** Si le jeu de données comporte plusieurs onglets, on identifie ici l'onglet où est disponible l'élément décrit.
- **Element_DisplayName:** nom de l'élément du jeu données qui vient d'être décrit.
- **Description:** «Une définition brève et complète de l'élément, énoncée au singulier, pouvant se distinguer des définitions d'autres éléments» (USDA, 2016). Il est important que les descriptions soient significatives en évitant que le texte ne contienne aucune information (Fig. 6.2).

B	C
Element_DisplayName	Description
number	Invoice number
date	Invoice date
status	Invoice status
amount	Invoice amount
customer_no	Customer number

B	C
Element_DisplayName	Description
number	Invoice autogenerated number, starting from 1 each year. Number is generated when invoice gets approved.
date	Invoice issued date. Null for working copy invoices. Automatically set to today's date on invoice approval.
status	Invoice status. 'W' - working copy, 'A' - approved invoice, 'C' - cancelled.
amount	Invoice net amount in USD
customer_no	Number of customer invoice was issued to. Ref: customers.

Figure 6.2. *Descriptions contenant zéro information à gauche et des descriptions contenant des informations utiles à droite (Source: Kononow, 2017).*

- **Unit:** unité de mesure adoptée pour les éléments.
- **Data_Type:** type des valeurs de données contenues dans le champ (par exemple, entier, date, etc.).
- **Character_Length:** La longueur des valeurs de données contenues dans le champ. «Par exemple, la longueur maximale d'Excel est de 255, indiquez donc 255 ou moins» (USDA, 2016).
- **Acceptable_Values:** la liste des valeurs acceptables dans ce champ. Les symboles adoptés pour séparer les valeurs "|", la plage de cellules "[a, b]", etc., sont basés sur les normes ISO (ISO 80000-2: 2009, 2009).
- **Required:** indiquez l'exigence de valeurs dans le champ pour le statut et la validité du jeu de données. Il est indiqué par y (oui) ou n (non). Si oui, les valeurs NULL ne sont pas acceptées pour ce champ dans le jeu de données.
- **Accepts_NullValue:** exprime la possibilité de valeurs nulles dans le champ correspondant du jeu de données. Cela nécessitait l'exécution de calculs sur les données. Il est indiqué par y (oui) ou n (non). Si oui, les valeurs nulles sont acceptées pour ce champ du jeu de données.

La «description des éléments du jeu de données» est généralement basée sur l'étude des variables du jeu de données (en-têtes de colonnes), ce qui permet de consulter leur signification et les éléments qu'ils contiennent. Ainsi, en utilisant cette structure, «Element_DisplayName» correspond aux noms des en-têtes de colonne. Ainsi, comme le montre la figure 6.3, s'il y a 4 colonnes dans un onglet, le dictionnaire du jeu de données aura au moins 4 lignes correspondant aux 4 colonnes de l'onglet.

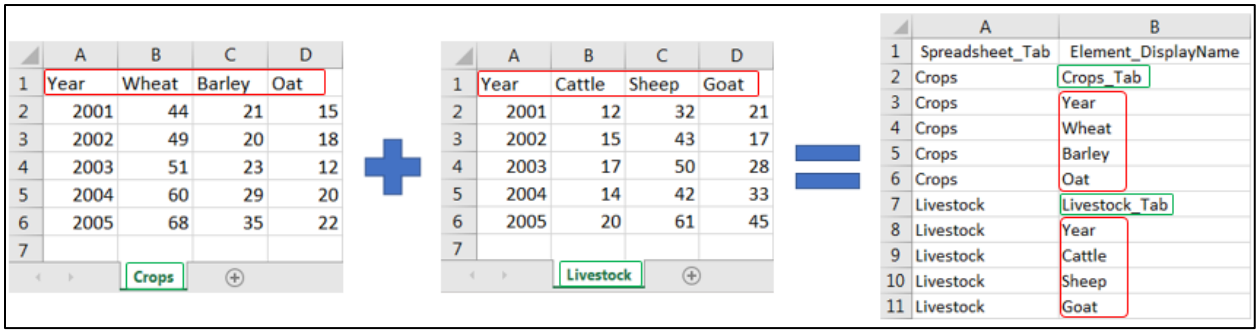
	A	B	C	D
1	Year	Wheat	Barley	Oat
2	2001	44	21	15
3	2002	49	20	18
4	2003	51	23	12
5	2004	60	29	20
6	2005	68	35	22
7				

Crops

	A	B
1	Spreadsheet_Tab	Element_DisplayName
2	Crops	Year
3	Crops	Wheat
4	Crops	Barley
5	Crops	Oat
6		

Figure 6.3. *Arrangements des en-têtes de colonne dans la feuille de calcul «DataDictionary_ElementsDescriptions».*

Lorsque le jeu de données est composé de plusieurs fichiers (et onglets), les différents éléments peuvent tous être répertoriés dans le même fichier «Description des éléments du jeu de données». Dans ce cas, il est bon d'ajouter une ligne avec la description de l'onglet pour chaque fichier de jeu de données (Fig. 6.4).



	A	B	C	D
1	Year	Wheat	Barley	Oat
2	2001	44	21	15
3	2002	49	20	18
4	2003	51	23	12
5	2004	60	29	20
6	2005	68	35	22
7				

	A	B	C	D
1	Year	Cattle	Sheep	Goat
2	2001	12	32	21
3	2002	15	43	17
4	2003	17	50	28
5	2004	14	42	33
6	2005	20	61	45
7				

	A	B
1	Spreadsheet_Tab	Element_DisplayName
2	Crops	Crops_Tab
3	Crops	Year
4	Crops	Wheat
5	Crops	Barley
6	Crops	Oat
7	Livestock	Livestock_Tab
8	Livestock	Year
9	Livestock	Cattle
10	Livestock	Sheep
11	Livestock	Goat

Figure 6.4. *Disposition des en-têtes de colonnes et des lignes de description des onglets dans la feuille de calcul «DataDictionary_ElementsDescriptions».*

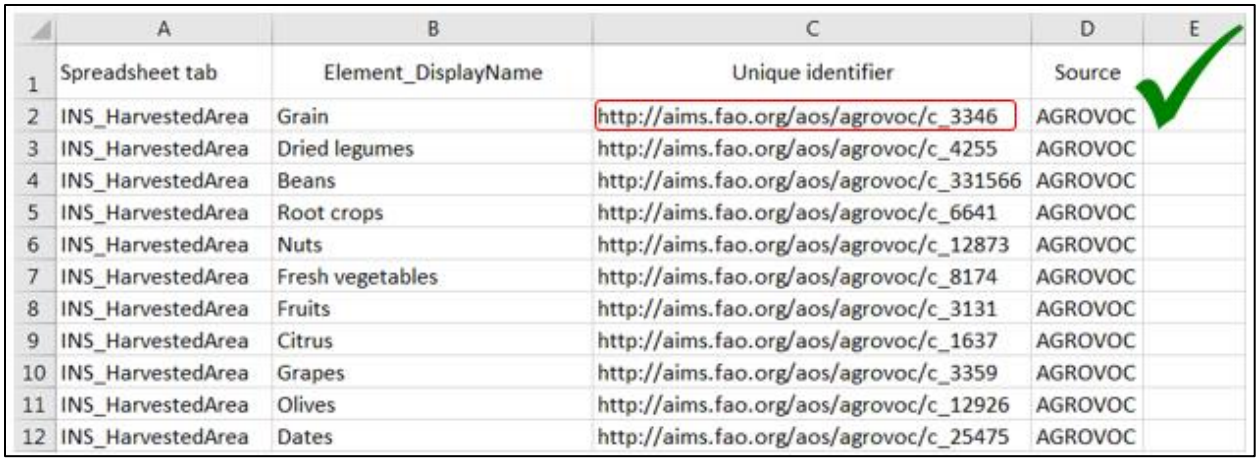
Un exemple complet de «Description des éléments du jeu de données» est présenté dans l'Annexe A (Fig. A.2, Fig. A.3 et Fig. A.4).

6.3. Identifiant unique

Pour vous assurer de résoudre toute ambiguïté possible, le lien correspondant aux termes et concepts du jeu de données vers le thésaurus multilingue en ligne est indiqué dans l'onglet identifiant unique. Ceci est très utile pour éviter tout malentendu sur les éléments analysés et rapportés dans l'ensemble de données (espèces de plantes, animaux, etc.).

Ce fichier est structuré des champs suivants (Fig. 6.5):

- **Spreadsheet_Tab:** Si la feuille de calcul a plusieurs onglets, on identifie ici l'onglet où est disponible l'élément décrit.
- **Element_DisplayName:** nom identifié de l'élément du jeu de données. Tous les éléments du jeu de données (valeurs, titres, etc.) peuvent être identifiés pour résoudre toute ambiguïté possible.
- **Unique_Identifier:** Indique le lien de référence ou, de préférence, les URI du thésaurus multilingue en ligne. Un identifiant de ressource uniforme (URI) est un identifiant unique qui rend le contenu adressable sur Internet en ciblant des éléments de manière unique (Rouse, 2014).
- **Source:** nom du thésaurus en ligne adopté pour identifier les adresses URI ou autres liens de référence (par exemple, AGROVOC, USDA, etc.).



	A	B	C	D	E
1	Spreadsheet tab	Element_DisplayName	Unique identifier	Source	
2	INS_HarvestedArea	Grain	http://aims.fao.org/aos/agrovoc/c_3346	AGROVOC	
3	INS_HarvestedArea	Dried legumes	http://aims.fao.org/aos/agrovoc/c_4255	AGROVOC	
4	INS_HarvestedArea	Beans	http://aims.fao.org/aos/agrovoc/c_331566	AGROVOC	
5	INS_HarvestedArea	Root crops	http://aims.fao.org/aos/agrovoc/c_6641	AGROVOC	
6	INS_HarvestedArea	Nuts	http://aims.fao.org/aos/agrovoc/c_12873	AGROVOC	
7	INS_HarvestedArea	Fresh vegetables	http://aims.fao.org/aos/agrovoc/c_8174	AGROVOC	
8	INS_HarvestedArea	Fruits	http://aims.fao.org/aos/agrovoc/c_3131	AGROVOC	
9	INS_HarvestedArea	Citrus	http://aims.fao.org/aos/agrovoc/c_1637	AGROVOC	
10	INS_HarvestedArea	Grapes	http://aims.fao.org/aos/agrovoc/c_3359	AGROVOC	
11	INS_HarvestedArea	Olives	http://aims.fao.org/aos/agrovoc/c_12926	AGROVOC	
12	INS_HarvestedArea	Dates	http://aims.fao.org/aos/agrovoc/c_25475	AGROVOC	

Figure 6.5. *Représentation d'un fichier standard «DataDictionary_UniqueIdentifier». En cas de besoin, la colonne "Notes" peut être ajoutée.*

7. Conclusions et Recommandations

Les travaux de curation de jeu de données peuvent être difficiles en fonction de l'état des données. En général, ils reposent sur une normalisation du contenu du jeu de données et sur la création de la documentation nécessaire. En suivant les étapes mentionnées précédemment, à partir d'un fichier de jeu de données au format Excel, nous terminons avec plusieurs fichiers au format CSV. Ainsi, le résultat de ce travail est un dossier pouvant être compressé, contenant les fichiers de dictionnaires de données et ceux du jeu de données au format CSV. Même d'autres éléments (par exemple, des images pertinentes du site) peuvent y être ajoutés afin de conserver tout le matériel du jeu de données dans un seul élément.

Cependant, il peut être difficile et coûteux, en termes de temps et d'argent, de traiter et de prendre en charge les processus de curation des jeux de données lorsque les données ont été collectées, gérées et analysées longtemps auparavant. Pour cette raison, la plus grande recommandation est de prendre en compte tous ces aspects dès les premières étapes de la collecte des données ; afin que ces pratiques deviennent une partie intégrante du travail de l'auteur, réduisant ainsi la lourdeur de cette activité et garantissant de meilleurs résultats.

Références

Bahlai, C., & Teal, T., (Eds). (2017, April). *Data Carpentry: Data Organization in Spreadsheets Ecology lesson* (Version 2017.04.0). Retrieved from http://www.datacarpentry.org/spreadsheet-ecology-lesson/

Big Data Platform Metadata Working Group, CGIAR. (2019). *CG Core metadata reference guide*. Retrieved from https://agriculturalsemantics.github.io/cg-core/cgcore.html

Briny, K. [University of Wisconsin Data Services]. (2015, January 23). *Data Management: Data Dictionaries*. Retrieved from https://www.youtube.com/watch?v=Fe3i9qyqPjo

Callahan, J. (2009). *Standard Latitudes and Longitudes*. Retrieved from http://mazamascience.com/WorkingWithData/?p=103

CGIAR Research Program on Grain Legumes and Dryland Cereals (GLDC), International Center for Agricultural Research in Dry Areas (ICARDA), CGIAR Research Program on Roots, Tubers and Bananas (RTB), WorldFish, CodeObia. (2019). *Monitoring, Evaluation & Learning (MEL): User Guide*. Retrieved July 17, 2019, from https://cgiarmel.atlassian.net/wiki/x/GoCC

Hodge, A. (2015). *File Naming Best Practices*. Retrieved from https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-naming

International Institute of Tropical Agriculture (IITA). (2019). *IITA Data Curation Guide*. Communication Unit, Data Management Section: Ibadan, Nigeria.

International Organization for Standardization. (2009). *ISO 80000-2:2009*. Retrieved from https://www.iso.org/standard/31887.html

Khawam, H., & Najjar, D. (2017). *Statistics on Gender and Education in Tunisia.* Retrieved from https://hdl.handle.net/20.500.11766.1/7MMLXI

Kononow, P. (2017, August 29). *Captain Obvious' Guide to Column Descriptions - Data Dictionary Best Practices* [Blog post]. Retrieved from https://dataedo.com/blog/captain-obivous-guide-to-column-descriptions-data-dictionary-best-practices

Munoz, T., Flanders, J., Senseney, M., Davis, R., Hsu, P.H., Little, J., Jackson, L.S., Renear, A., & Trainor, K. (2017). *DH Curation Guide: a community resource guide to data curation in the digital humanities* (FAQ). Retrieved December 5, 2018, from https://guide.dhcuration.org/about/

Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC). (2018). Data Management. Retrieved December 5, 2018, from https://daac.ornl.gov/datamanagement/

Rouse, M., & Wigmore, I. (2014). *Definition: unique identifier (UID)*. Retrieved from https://internetofthingsagenda.techtarget.com/definition/unique-identifier-UID

Ruggles, S. (2018). The Importance of Data Curation. In Vannette, D., Krosnick, J. (Eds). *The Palgrave Handbook of Survey Research* (pp. 303-308). Palgrave Macmillan: Cham. Retrieved from https://doi.org/10.1007/978-3-319-54395-6_39

United States Department of Agriculture (USDA) (2016). *Ag Data Commons Data Submission Manual v1.3*. National Agricultural Library. Retrieved from https://data.nal.usda.gov/book/export/html/2769

United States Department of Agriculture (USDA). [National Agricultural Library]. (2017, August 9). *ADC 18 - Convert data files to CSV format*. Retrieved from https://www.youtube.com/watch?v=szDWlvQOa_g&index=19&list=PL_8uALA03ZsWQ44QNKo4_ZSYSQP7gJ9h7

White, E.P., Baldridge, E., Brym, Z.T., Locey, K.J., McGlinn, D.J., & Supp, S.R. (2013). *Nine simple ways to make it easier to (re)use your data*. Retrieved from https://doi.org/10.7287/peerj.preprints.7v2

Wikipedia contributors. (2018). Comma-separated values. In *Wikipedia, The Free Encyclopedia*. Retrieved December 5, 2018, from https://en.wikipedia.org/w/index.php?title=Comma-separated_values&oldid=922764682

Wikipedia contributors. (2019). Handle System. In *Wikipedia, The Free Encyclopedia*. Retrieved October 29, 2019, from https://en.wikipedia.org/w/index.php?title=Handle_System&oldid=923416074

Zwicker, S., in Hsu, L. (2016, December 7). *How "clean" should an Excel file be to be considered machine readable*. Retrieved from https://my.usgs.gov/confluence/pages/viewpage.action?pageId=559852026

The Dataverse Project. (2019). *About the Project*. Retrieved from https://dataverse.org/about

Annexe A

	A	B	C	D	E	F	G
1	Spredsheet_Tab	Site_ID	Location	Country_ISO3	Latitudes	Longitudes	Altitude
2	PlantCover	PL_07	Tataouine	TUN	32.94892	10.48665	247
3	PlantCover	PL_12	Ifrane	MAR	33.30687	-5.01723	1693
4							

Figure A.1. *Exemple de fichier «Site_Summary». Veuillez noter que chaque fois qu'un nouveau fichier est développé, les descriptions et les détails de l'en-tête de colonne doivent être consignés dans l'onglet "DataDictionary_ElementsDescriptions".*

	A	B	C
1	Years	NetPrimaryFemale	NetPrimaryMale
2	2000	94.15018	96.6708
3	2001	96.45512	98.07967
4	2002	NA	NA
5	2003	NA	NA
6	2004	NA	NA
7	2005	98.6595	99.06998
8	2006	98.66384	98.59785
9	2007	97.41568	97.87909
10	2008	97.60895	98.38915
11	2009	98.26036	98.9663
12	2010	NA	NA
13	2011	NA	NA
14	2012	NA	NA
15	2013	NA	NA
16	2014	NA	NA

WorldBank_Education

	A	B	C
1	Years	BasicSecondary_Male	BasicSecondary_Female
2	2000	469202	493783
3	2001	497945	529867
4	2002	507290	549943
5	2003	511999	564239
6	2004	512001	572877
7	2005	505330	570187
8	2006	511128	577688
9	2007	500517	569068
10	2008	467328	538815
11	2009	447369	520339
12	2010	433814	502584
13	2011	428109	494349
14	2012	418498	490102
15	2013	408292	479153
16	2014	402896	473815

INS_StudentsEducation

Figure A.2 (À Gauche) and A.3 (À droite). *Image de deux jeux de données organisés. Jeux de données WorldBank_Education à gauche et INS_StudentsEducation à droite.*

	A	B	C	D	E	F	G	H	I
1	Spreadsheet_Tab	Element_DisplayName	Description	Unit	Data_Type	Character_Length	Acceptable_Values	Required	Accepts_NullValue
2	WorldBank_Education	World Bank_Education_Tab	Data about education participation in Tunisia from 1999 to 2014. Source: World Development Indicators, THE WORLD BANK. Last update 1/2/2017. Sheet: Series: World bank indicators file. Related link: http://data.worldbank.org/data-catalog/world-development-indicators	NA	NA	NA	NA	NA	NA
3	WorldBank_Education	Years	The year to which this analysis refers.	yyyy	date	4	2000|2014	y	n
4	WorldBank_Education	NetPrimaryFemale	The element full name is "Adjusted net enrollment rate, primary, female (% of primary school age children)". Adjusted net enrollment is the number of pupils of the school-age group for primary education, enrolled either in primary or secondary education, expressed as a percentage of the total population in that age group.	%	decimals	255	NA	n	y
5	WorldBank_Education	NetPrimaryMale	The element full name is "Adjusted net enrollment rate, primary, male (% of primary school age children)". Adjusted net enrollment is the number of pupils of the school-age group for primary education, enrolled either in primary or secondary education, expressed as a percentage of the total population in that age group.	%	decimals	255	NA	n	y
6	INS_StudentsEdu	INS_StudentsEdu_Tab	Data about male and female basic and secondary education in Tunisia. Last update: 17/03/2016. Source: Ministry of education (Statistique Tunisia). Related link: http://www.ins.tn/en/themes/education	NA	NA	NA	NA	NA	NA
7	INS_StudentsEdu	Years	The year to which this analysis refers.	yyyy	date	4	2000|2014	y	n
8	INS_StudentsEdu	BasicSecondary_Male	The element full name is: "Number of male students in the second cycle of basic education and secondary public education". It corresponds to the male students registered in Tunisia for the different years	Individuals	numeric	6	0|∞	n	y
9	INS_StudentsEdu	BasicSecondary_Female	The element full name is "Number of female students in the second cycle of basic education and secondary public education".It corresponds to the female students registered in Tunisia for the different years	Individuals	numeric	6	0|∞	n	y

Figure A.4. *«DataDictionary_ElementsDescriptions» des deux jeux de données organisés présentés dans les images A.2 et A.3. Source: Khawam, 2017.*

Annexe B - Tools and Weblinks

Resource name: AGROVOC
Type: Vocabulary
Description: Online multilingual thesaurus to find URIs for the unique identifier field.
Link: http://aims.fao.org/standards/agrovoc/functionalities/search

Resource name: USDA Thesaurus
Type: Vocabulary
Description: Online multilingual thesaurus to find URIs for the unique identifier field.
Link: https://agclass.nal.usda.gov/mtwdk.exe

Resource name: Catalogue of life
Type: Species global index
Description: Species global index where to find URNs for the unique identifier field.
Link: http://www.catalogueoflife.org/col/search/all

Resource name: The International Plant Name Index
Type: Plant index
Description: Index where to check and find plant names and associated bibliography details.
Link: http://www.ipni.org/ipni/plantnamesearchpage.do

Resource name: The Plant List
Type: Plant index
Description: Index where to check and find plant names and associated bibliography details.
Link: http://www.theplantlist.org/

Resource name: The International Union for Conservation of Nature's Red List of Threatened Species
Type: Species global index
Description: Information source on the biodiversity and the global conservation status of animal, fungi and plant species.
Link: https://www.iucnredlist.org/

Resource name: USDA Ag Data Commons Beta
Type: Dataset Repository
Description: USDA Dataset Repository of curated datasets.
Link: https://data.nal.usda.gov/

Resource name: Converting from CSV to Excel worksheet
Type: Video
Description: Video tutorial on how to convert CSV files to Excel format.
Link: https://www.youtube.com/watch?v=wpDq96Y_wgw

Resource name: How to open a CSV file in Excel?
Type: Guide
Description: Short guide on how to convert CSV files to Excel format.
Link: https://www.copytrans.net/support/how-to-open-a-csv-file-in-excel/

Resource name: How to Convert Delimited Text Files to Excel Spreadsheets
Type: Guide
Description: Short guide on how to convert CSV files to Excel format.
Link: https://www.makeuseof.com/tag/how-to-convert-delimited-text-files-into-excel-spreadsheets/

Resource name: Saving an Excel File as a CSV File
Type: Guide
Description: Short guide on how to convert Excel file to CSV format.
Link: https://knowledgebase.constantcontact.com/articles/KnowledgeBase/6409-saving-an-excel-file-as-a-csv-file?lang=en_US

Resource name: 10 Super Neat Ways to Clean Data in Excel Spreadsheets
Type: Guide
Description: Short guide and video tutorial about simple ways to facilitate the data cleaning in Excel spreadsheet.
Link: https://trumpexcel.com/clean-data-in-excel/

Resource name: Talend Data Preparation
Type: Tool
Description: Automatic tool to support for cleaning and formatting actions.
Link: https://www.talend.com/products/data-preparation/

Resource name: Data types in Microsoft Excel
Type: Guide
Description: Short guide on how to identify the data type in Excel files.
Link: https://www.promotic.eu/en/pmdoc/Subsystems/Db/Excel/DataTypes.htm

Resource name: Data types used by Excel
Type: Guide
Description: Short guide on how to identify the data type in Excel files.
Link: https://docs.microsoft.com/en-us/office/client-developer/excel/data-types-used-by-excel

Resource name: USGS Data Management
Type: Guide
Description: Guide on data dictionary creation and management.
Link: https://www.usgs.gov/products/data-and-tools/data-management/data-dictionaries

Resource name: Creating machine readable data
Type: Guide
Description: Guide about releasing data in formats that are machine-readable and allow for easy reuse under the Western Australian Whole of Government Open Data Policy.
Link: https://data.wa.gov.au/junk/fact-sheets-and-toolkit/creating-machine-readable-data

Resource name: Excel specifications and limits
Type: Factsheet
Description: Excel Worksheet and workbook specifications and limits.
Link: https://support.office.com/en-us/article/excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3

Resource name: Coordinates conversion
Type: Tool
Description: Conversions of latitude and longitude geographic coordinates in different formats.
Link: https://www.sunearthtools.com/dp/tools/conversion.php?lang=en

Resource name: Mathematical signs and symbols
Type: Encyclopedia
Description: Mathematical signs and symbols definitions based on the international standards (ISO 31-11:1992).
Link: https://en.wikipedia.org/wiki/ISO_31-11

Resource name: The Relevance of Rest Periods in Rangeland Management for Plant Density in Tataouine, Tunisia, Spring 2017
Type: Dataset
Description: Examples of ICARDA's curated dataset in CSV format.
Link: https://hdl.handle.net/20.500.11766.1/FK2/X9IYIU

Resource name: Soil Moisture Records for Different Water Harvesting Treatments, Jordan, 2016/2017
Type: Dataset
Description: Examples of ICARDA's curated dataset in CSV format.
Link: https://hdl.handle.net/20.500.11766.1/7DYLFW