Importance et Principes De la Curation des Jeux de Données

Pietro Bartolini, Asma Jeitani

Monitoring Evaluation and Learning, Data Management and Geo-informatics Option by Context – Learning Week
1-7 November, Tunis, Tunisia

Curation des données : Pourquoi?

Les données sont la colonne vertébrale de toute recherche scientifique.

Presque toutes les activités de recherche produisent une sorte de jeux de données en tant que sous-produit.

Malheureusement, un jeu de données est souvent considéré comme moins important qu'un document ou un rapport: nous l'utilisons, l'organisant en fonction de nos besoins temporaires, puis nous le laissons dans un référentiel sans autre préparation.

Sans une curation appropriée, un jeu de données risque d'être totalement inutilisable pour les futurs projets de recherche et il est impossible de diffuser facilement les résultats de la recherche.

Un jeu de données ‘organisé’ est un jeu de données réutilisable qui peut être rendu public en tant que produit indépendant.

Curation des données : Comment?

Caractéristiques principales du jeu de données ‘organisé’ :

- Fournir des informations pertinentes sur le contenu et le contexte de la recherche.
- Contenir que des données brutes, sans élaboration supplémentaire.
- Respecter les normes applicables aux documents lisibles à la machine.

Une description détaillée du processus de curation est disponible dans le *Guide General de la Curation des Jeux de Données* dans le lien ci-dessous.

https://hdl.handle.net/20.500.11766/9400

Étapes Préliminaires

Afin d'éviter les erreurs et les pertes de données, ne modifiez pas le jeu de données d'origine. Créez une nouvelle copie.

Améliorez le titre du nouveau jeu de données en fournissant un contexte.

Ancien Titre du Jeu de Données:	Rangeland Species Composition
Nouveau Titre du Jeu de Données:	Annual and Perennial Rangeland Plant Cover and Species Composition, Tatatouine, Tunisia, November 2018

Si les données proviennent d'une citation d'article primaire, utilisez la convention de dénomination "Données tirées de: titre de l'article" "(USDA, 2016).

Curation des données – Gestion des Elaborations

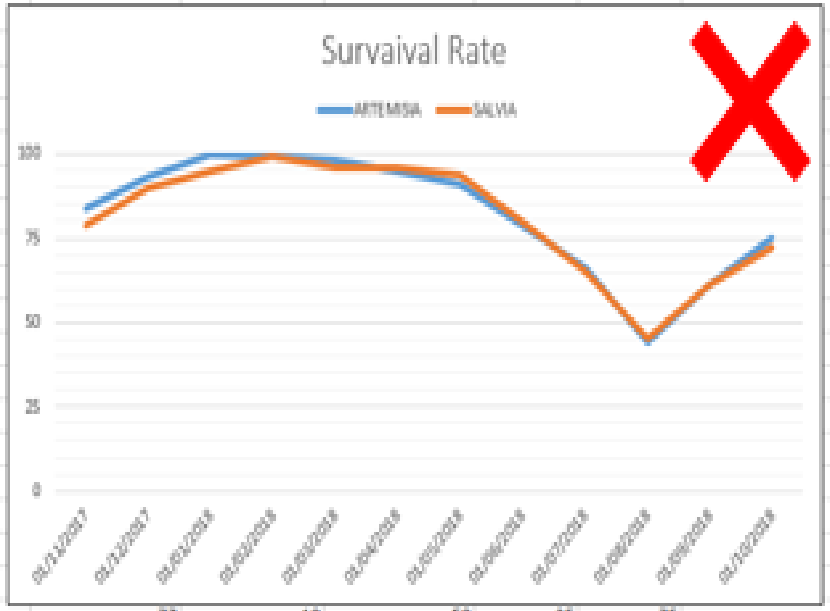
Le jeu de données final ne doit contenir que des données brutes, car chaque élaboration est sujette à erreur et d'autres chercheurs peut être voudrons faire une analyse différente.

- Pas de graphique
- Pas de formules
- Pas de pourcentages
- Pas d’élaboration

En fournissant uniquement des données brutes, nous permettons aux futurs utilisateurs de les utiliser pour leurs recherches, en évitant le risque de reproduire des erreurs d'élaboration.

Curation des données – Gestion des Elaborations

	A	B	C	D	E	F	G	H
1	Period	Species	Number survived	No of Dead	No of Living	% of Dead	% of Living	
2	20171115	ARTEMISIA	92	15	77	16	84	
3	20171115	SALVIA	112	24	88	21	79	
4	20171215	ARTEMISIA	116	8	108	7	93	
5	20171215	SALVIA	135	13	122	10	90	
6	20180115	ARTEMISIA						
7	20180115	SALVIA						
8	20180215	ARTEMISIA						
9	20180215	SALVIA						
10	20180315	ARTEMISIA						
11	20180315	SALVIA						
12	20180415	ARTEMISIA						
13	20180415	SALVIA						
14	20180515	ARTEMISIA						
15	20180515	SALVIA						
16	20180615	ARTEMISIA						
17	20180615	SALVIA						
18	20180715	ARTEMISIA						
19	20180715	SALVIA						
20	20180815	ARTEMISIA						
21	20180815	SALVIA						
22	20180915	ARTEMISIA						
23	20180915	SALVIA						
24	20181015	ARTEMISIA	77	19	58	25	75	
25	20181015	SALVIA	99	28	71	28	72	

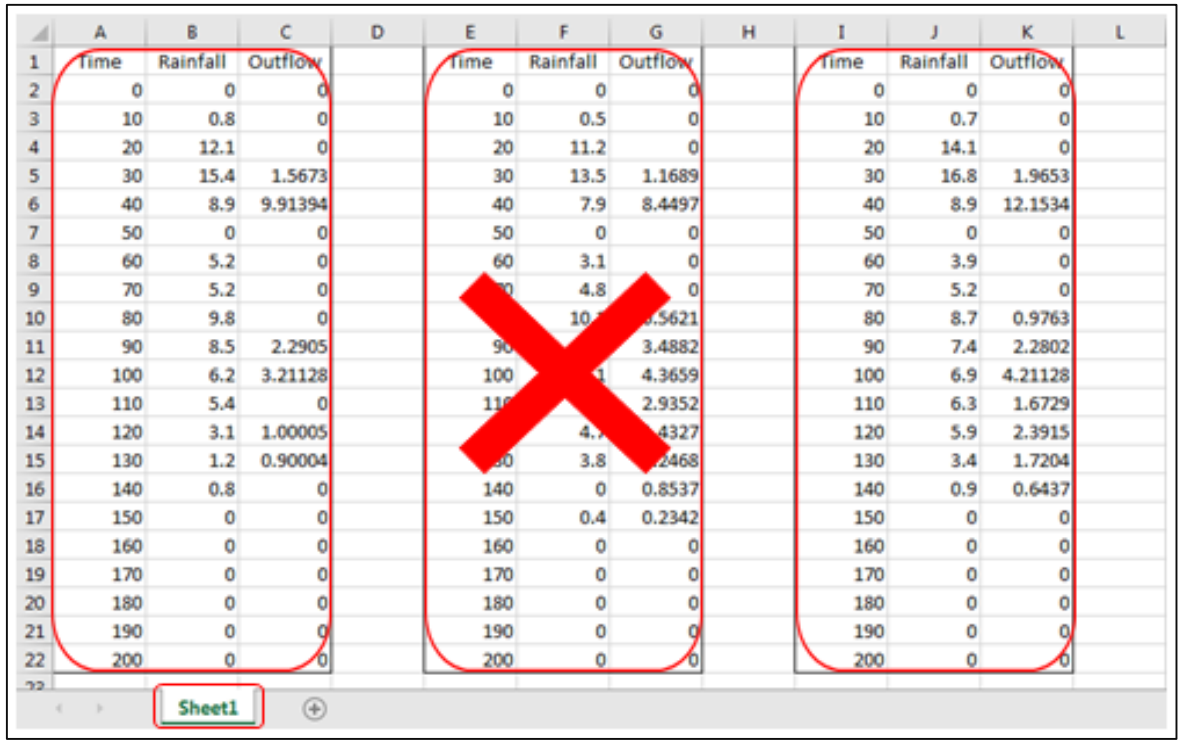


	A	B	C	D	E	F	G	H
1	Period	Species	Number survived	No of Dead	No of Living	% of Dead	% of Living	
2	20171115	ARTEMISIA	92	15	77	16	84	
3	20171115	SALVIA	112	24	88	21	79	
4	20171215	ARTEMISIA	116	8	108	7	93	
5	20171215	SALVIA	135	13	122	10	90	
6	20180115	ARTEMISIA	139	0	139	0	100	
7	20180115	SALVIA	152	8	144	5	95	
8	20180215	ARTEMISIA	137	1	136	1	99	
9	20180215	SALVIA	149	1	148	1	99	
10	20180315	ARTEMISIA	121	2	119	2	98	
11	20180315	SALVIA	126	5	121	4	96	
12	20180415	ARTEMISIA	101	5	96	5	95	
13	20180415	SALVIA	115	5	110	4	96	
14	20180515	ARTEMISIA	82	7	75	9	91	
15	20180515	SALVIA	108	7	101	6	94	
16	20180615	ARTEMISIA	78	16	62	21	79	
17	20180615	SALVIA	95	19	76	20	80	
18	20180715	ARTEMISIA	64	22	42	34	66	
19	20180715	SALVIA	83	29	54	35	65	
20	20180815	ARTEMISIA	48	27	21	56	44	
21	20180815	SALVIA	71	39	32	55	45	
22	20180915	ARTEMISIA	59	23	36	39	61	
23	20180915	SALVIA	80	31	49	39	61	
24	20181015	ARTEMISIA	77	19	58	25	75	
25	20181015	SALVIA	99	28	71	28	72	

Curation des données – Arrangement des tableaux

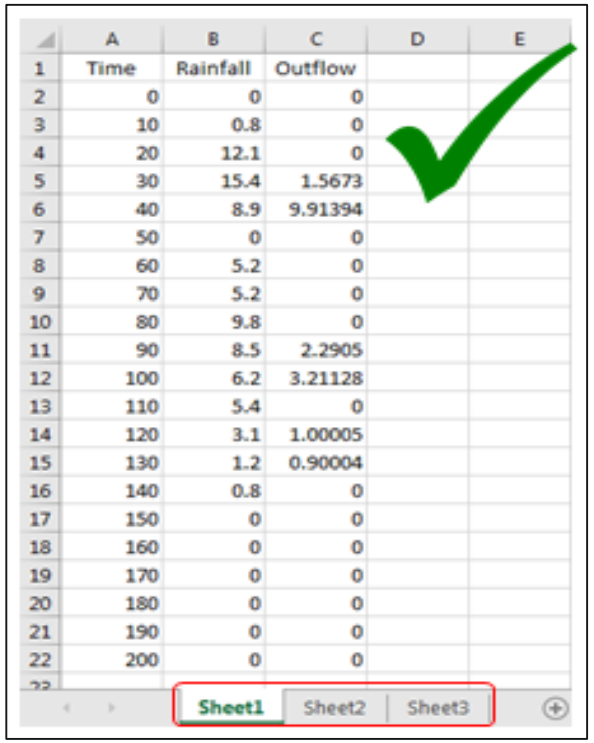
Pour être lisible par machine, le jeu de données ne peut pas contenir plusieurs tableaux dans une seule feuille de calcul, en utilisant des lignes ou des colonnes vides pour séparer les données.

Chaque feuille de calcul doit contenir un seul tableau.



	A	B	C	D	E	F	G	H	I	J	K	L
1	Time	Rainfall	Outflow		Time	Rainfall	Outflow		Time	Rainfall	Outflow	
2	0	0	0		0	0	0		0	0	0	
3	10	0.8	0		10	0.5	0		10	0.7	0	
4	20	12.1	0		20	11.2	0		20	14.1	0	
5	30	15.4	1.5673		30	13.5	1.1689		30	16.8	1.9653	
6	40	8.9	9.91394		40	7.9	8.4497		40	8.9	12.1534	
7	50	0	0		50	0	0		50	0	0	
8	60	5.2	0		60	3.1	0		60	3.9	0	
9	70	5.2	0		70	4.8	0		70	5.2	0	
10	80	9.8	0		80	10.1	0.5621		80	8.7	0.9763	
11	90	8.5	2.2905		90	7.4	3.4882		90	7.4	2.2802	
12	100	6.2	3.21128		100	5.1	4.3659		100	6.9	4.21128	
13	110	5.4	0		110	4.3	2.9352		110	6.3	1.6729	
14	120	3.1	1.00005		120	4.7	4327		120	5.9	2.3915	
15	130	1.2	0.90004		130	3.8	2468		130	3.4	1.7204	
16	140	0.8	0		140	0	0.8537		140	0.9	0.6437	
17	150	0	0		150	0.4	0.2342		150	0	0	
18	160	0	0		160	0	0		160	0	0	
19	170	0	0		170	0	0		170	0	0	
20	180	0	0		180	0	0		180	0	0	
21	190	0	0		190	0	0		190	0	0	
22	200	0	0		200	0	0		200	0	0	

Sheet1

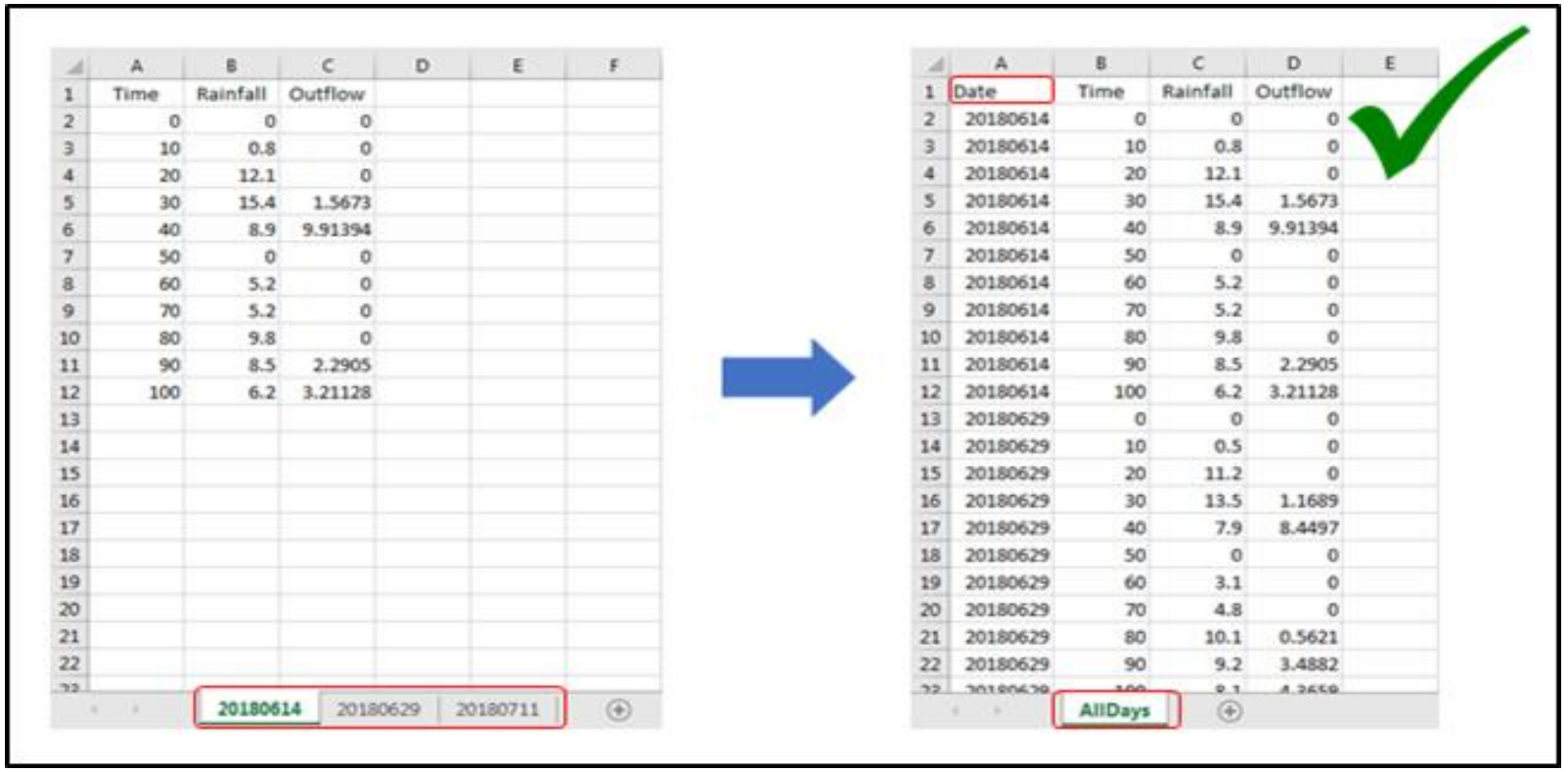


	A	B	C	D	E
1	Time	Rainfall	Outflow		
2	0	0	0		
3	10	0.8	0		
4	20	12.1	0		
5	30	15.4	1.5673		
6	40	8.9	9.91394		
7	50	0	0		
8	60	5.2	0		
9	70	5.2	0		
10	80	9.8	0		
11	90	8.5	2.2905		
12	100	6.2	3.21128		
13	110	5.4	0		
14	120	3.1	1.00005		
15	130	1.2	0.90004		
16	140	0.8	0		
17	150	0	0		
18	160	0	0		
19	170	0	0		
20	180	0	0		
21	190	0	0		
22	200	0	0		

Sheet1 Sheet2 Sheet3

Curation des données – Arrangement des tableaux

Si différentes feuilles de calcul contiennent des données similaires et partagent une structure similaire, il peut être nécessaire de les analyser ensemble. Afin de permettre au programme de reconnaitre la connexion, ils peuvent être fusionnés dans une seule feuille de calcul.



20180614

Time	Rainfall	Outflow
0	0	0
10	0.8	0
20	12.1	0
30	15.4	1.5673
40	8.9	9.91394
50	0	0
60	5.2	0
70	5.2	0
80	9.8	0
90	8.5	2.2905
100	6.2	3.21128

20180629

Time	Rainfall	Outflow
0	0	0
10	0.5	0
20	11.2	0
30	13.5	1.1689
40	7.9	8.4497
50	0	0
60	3.1	0
70	4.8	0
80	10.1	0.5621
90	9.2	3.4882

20180711

AllDays

Date	Time	Rainfall	Outflow
20180614	0	0	0
20180614	10	0.8	0
20180614	20	12.1	0
20180614	30	15.4	1.5673
20180614	40	8.9	9.91394
20180614	50	0	0
20180614	60	5.2	0
20180614	70	5.2	0
20180614	80	9.8	0
20180614	90	8.5	2.2905
20180614	100	6.2	3.21128
20180629	0	0	0
20180629	10	0.5	0
20180629	20	11.2	0
20180629	30	13.5	1.1689
20180629	40	7.9	8.4497
20180629	50	0	0
20180629	60	3.1	0
20180629	70	4.8	0
20180629	80	10.1	0.5621
20180629	90	9.2	3.4882

Remarque:
Lors de la fusion des feuilles de calcul, veillez à ajouter une colonne avec un code d'identification clair, tel que des dates ou des emplacements différents.

Curation des données – Gestion de données

La structure du jeu de données doit être propre et simple:

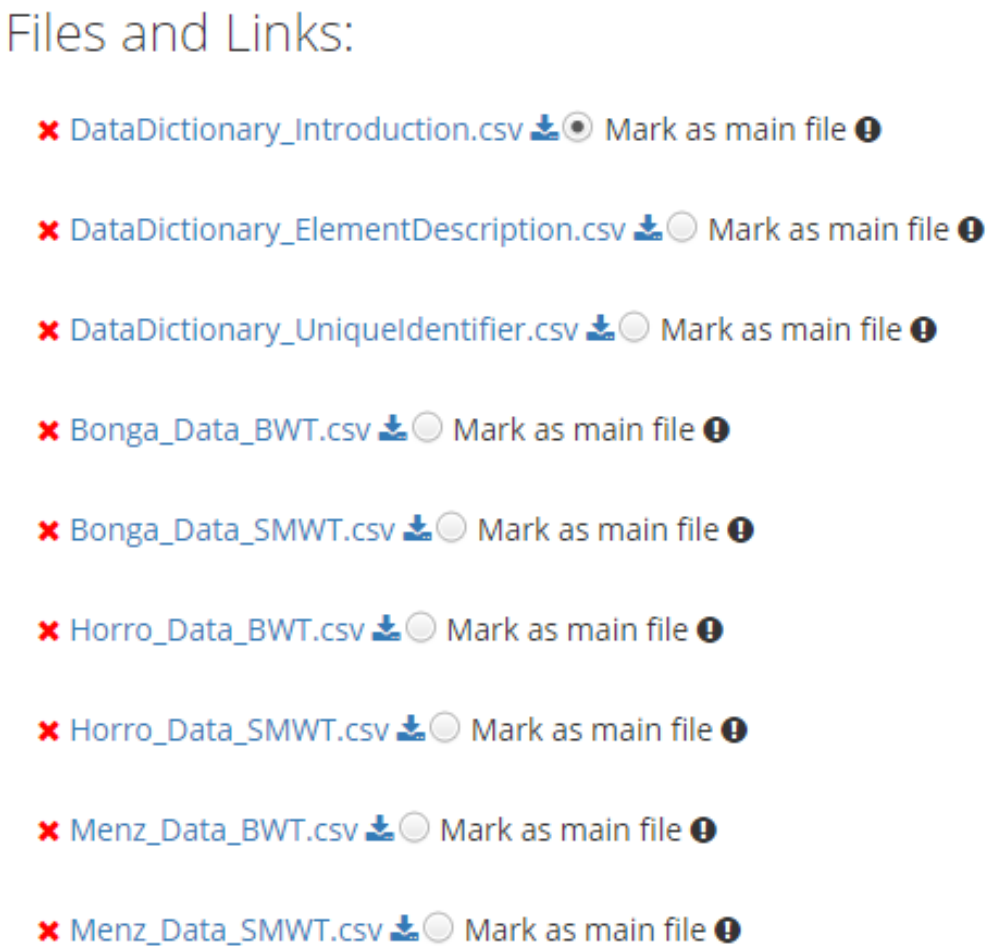
- Titre court des en-têtes de colonnes (pas d'espaces ni de symboles, écrivez dans un camel case ou utilisez un trait de soulignement)
- 1 seule information dans chaque cellule
- Aucune information vague ou trompeuse, utilisez si possible un code numérique
- Utilisez les normes ISO pour la date et l'heure (AAAAMMJJHHMM)
- Pas de cellules fusionnées
- Pas de commentaire, utilisez une colonne pour les notes
- Pas de cellules vides, utilisez NA pour les données manquantes ou nulles
- Pas de format de texte ou de couleur étrange (ils pourraient améliorer la lisibilité, en soulignant les données pertinentes, mais ils peuvent facilement être perdus lors du transfert, ce qui compromet la structure générale)

Format du Fichier

Quel que soit le logiciel utilisé, nous devons être assuré que les fichiers seront lisibles au futur, en utilisant différentes versions d'un logiciel sous licence ou sans licence.

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.

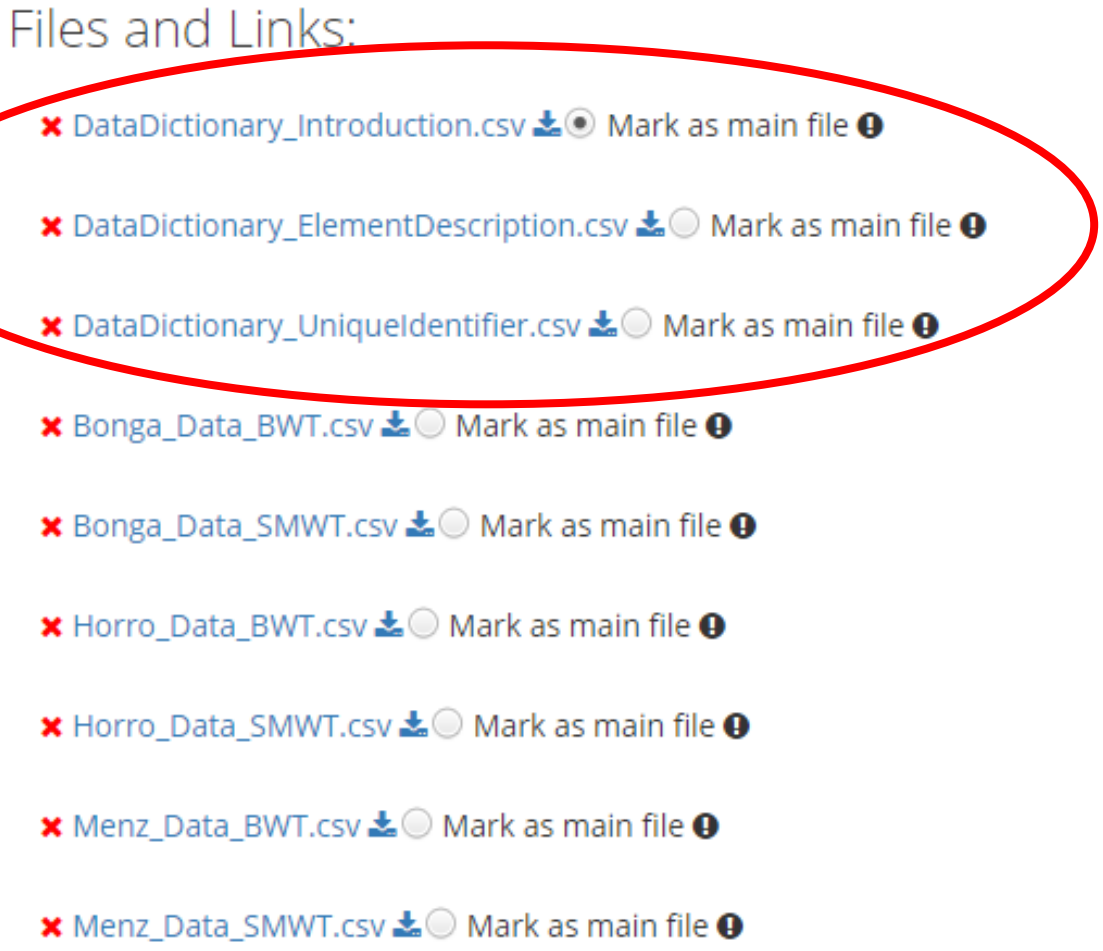
Comme le fichier CSV contient une seule feuille de calcul, un jeu de données téléversé dans MEL sera un ensemble de plusieurs fichiers CSV.



Fournir du Contexte - Dictionnaire de données

Afin de fournir des informations sur le contenu et le contexte de la recherche, chaque jeu de données doit inclure un dictionnaire de données complet:

- Dictionnaire de Données - Introduction au jeu de données
- Dictionnaire de Données - Description des Éléments du jeu de données
- Dictionnaire de Données - Identificateur unique



Introduction au jeu de données

L'introduction du jeu de données fournit une explication générale sur la portée et la création du jeu de données. L'introduction doit inclure:

- 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).
- 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.
- Author: premier auteur du jeu de données.
- CoAuthor: co-auteurs du jeu de données.

Introduction au jeu de données

Description	Summary	Start_Date	End_Date	Author	CoAuthor
Description libre de texte riche fournissant le plus d'explications possible sur le jeu de données.	Description plus courte du jeu de données, généralement pas plus d'une phrase ou deux.	AAAAMMJJ	AAAAMMJJ	Premier auteur du jeu de données.	Co-auteurs du jeu de données.

C'est la structure de base, mais l'onglet est personnalisable en fonction des besoins de l'auteur (c'est-à-dire l'ajout de sections spécifiques sur la localisation, les méthodologies et les notes complémentaires).

Description des éléments du jeu de données

Le composant le plus important du dictionnaire de données: fournit des explications sur la signification de chaque variable et les correspondances pour tout code utilisé.

Spreadsheet_Tab	Element_DisplayName	Description	Units	Data_Type	Character_ Length	Acceptable_ Values	Required	Accepts_ NullValue
Spreadsheet_Name	Spreadsheet_Name	Description of the spreadsheet content	NA	NA	NA	NA	NA	NA
Spreadsheet_Name	Variable_N1	Description of the variable meaning	Kg	Numeric	255	[x, z]	Y/N	Y/N
Spreadsheet_Name	Variable_N2	Description of the variable meaning	NA	Numeric	2	x|y|z	Y/N	Y/N
Spreadsheet_Name	Variable_N3	Description of the variable meaning	NA	Text	255	NA	Y/N	Y/N
Spreadsheet_Name	Variable_N3	Description of the variable meaning	YYYYMMDD	Date	8	[yyyymmdd, YYYYMMDD]	Y/N	Y/N

Description des éléments du jeu de données

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

- Spreadsheet_Tab: l'onglet où est disponible l'élément décrit.
- Element_DisplayName: nom de l'élément du jeu de données
- Description: Une définition brève et complète de l'élément pouvant se distinguer des définitions d'autres éléments

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.

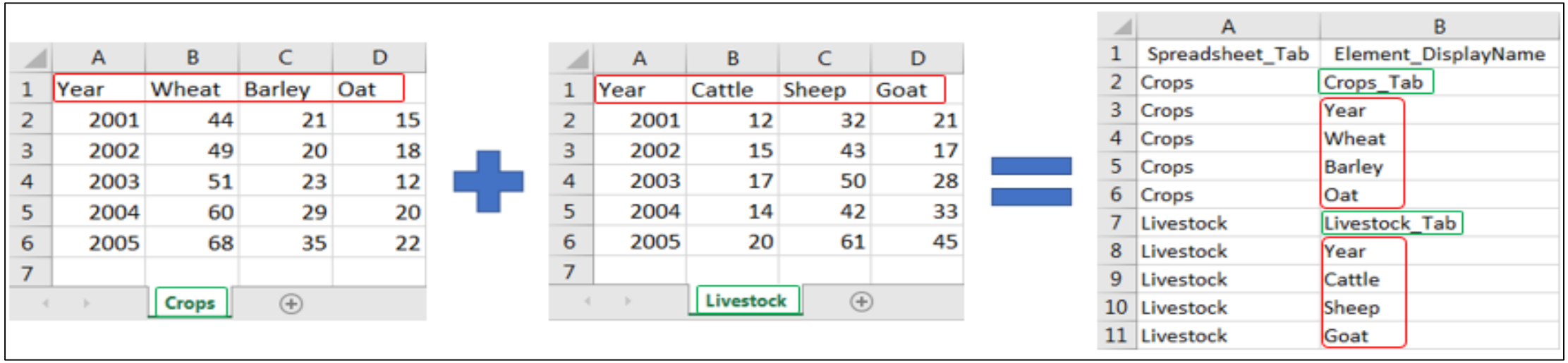
Description des éléments du jeu de données

- 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: longueur des valeurs de données contenues dans le champ (longueur maximal dans Excel est 255).
- Acceptable_Values: la liste des valeurs acceptables dans ce champ. Dans certains cas ça peut être des plages de cellules.

Acceptable_Values	
1|2|3|4|5	Sélection fixe parmi plusieurs options, séparées par une barre verticale.
Black|Red|White	
[0, 2000]	Plage de valeurs aléatoire. La plus petite valeur et la plus grande entre crochets, séparées par une virgule.

- Required: indique l'exigence de valeurs dans le champ pour le statut et la validité du jeu de données.
- Accepts_NullValue: exprime la possibilité de valeurs nulles dans le champ correspondant du jeu de données.

Description des éléments du jeu de données



	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	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				

Livestock

=

	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

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 (Bonechi 2018).

Identificateur Unique

Nous supposons souvent que l'utilisation de certains termes est claire, mais ce n'est pas toujours le cas, surtout en dehors de notre équipe de recherche.

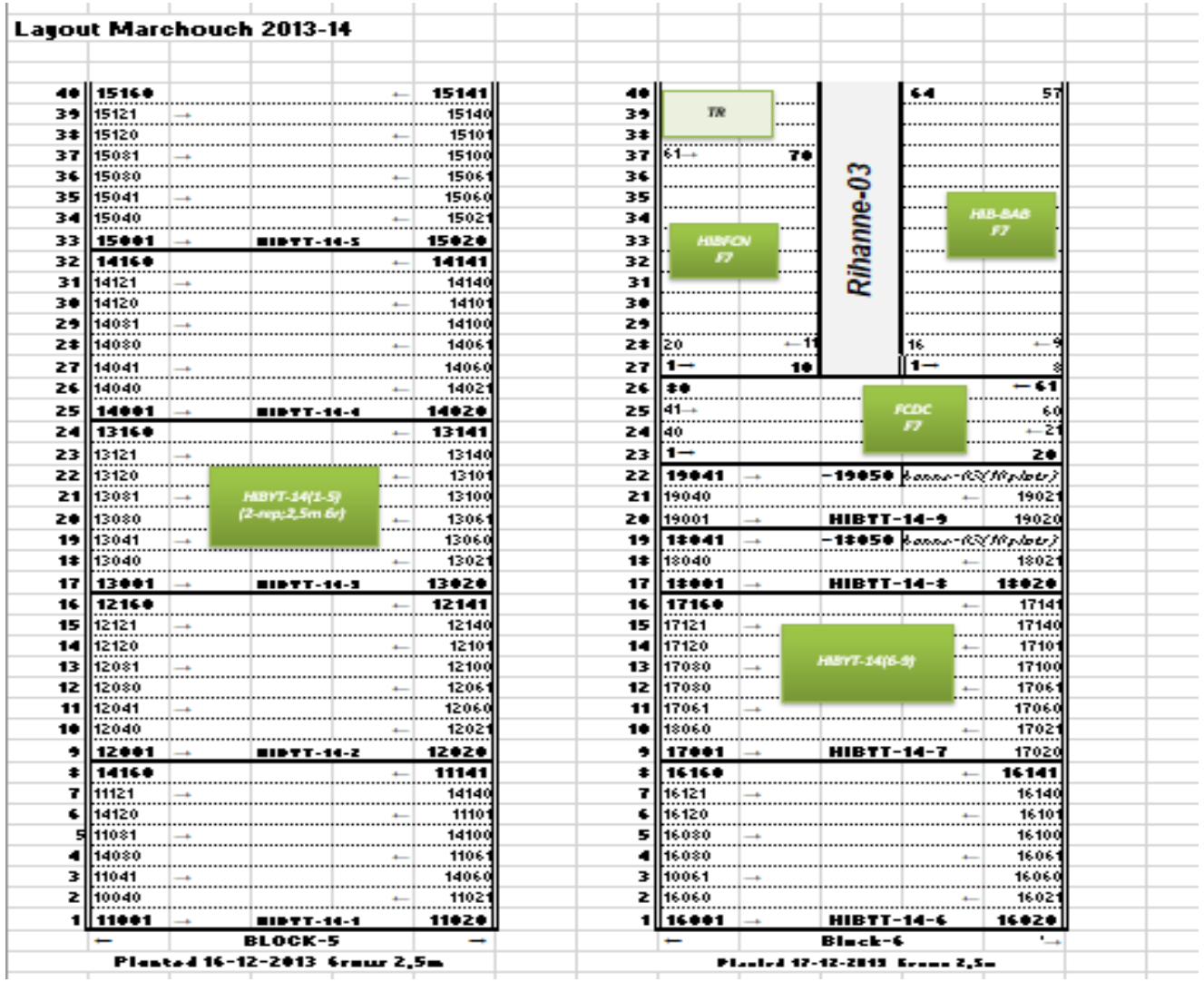
Pour vous assurer de résoudre toute ambiguïté possible, le lien correspondant aux termes et concepts du jeu de données vers le thésaurus multilingue en ligne est indiqué dans l'onglet identifiant unique. Ceci est très utile pour éviter tout malentendu sur les éléments analysés et rapportés dans l'ensemble de données (espèces de plantes, animaux, etc.).
(Bonechi 2018).

Spreadsheet_Tab	Element_DisplayName	Unique_Identifier	Source
Spreadsheet_Name1	Earth dams	http://aims.fao.org/aos/agrovoc/c_32435	AGROVOC
Spreadsheet_Name2	Sheep fattening	http://lod.nal.usda.gov/nalt/92111	USDA
Spreadsheet_Name3	Barley	http://aims.fao.org/aos/agrovoc/c_823	AGROVOC
Spreadsheet_Name3	Malting Barley	http://aims.fao.org/aos/agrovoc/c_25485	AGROVOC

Recommandation finale

«Le Guide General de la Curation des Jeux de Données» établit des normes, mais c'est toujours un travail en cours. Certains types de fichiers peuvent nécessiter des solutions ad hoc (par exemple, des données de disposition de parcelles, ou de gros fichiers sur des données génétiques, en utilisant un format de fichier différent).

Après l'exercice d'aujourd'hui, chacun d'entre vous peut commencer à l'appliquer jusqu'à ce que cela devienne une pratique courante. Vous pouvez également signaler les problèmes potentiels au personnel de curation des données, ce qui vous permet d'élaborer davantage la future version du guide.



Merci pour votre attention!