ICRISAT
INTERNATIONAL CROPS RESEARCH
INSTITUTE FOR THE SEMI-ARID TROPICS

ICARDA
Science for resilient livelihoods in dry areas

World Agroforestry

ILRI
INTERNATIONAL
LIVESTOCK RESEARCH
INSTITUTE

IFAD
Investing in rural people

The Dataverse Project

A brief introduction on data curation process

Pietro Bartolini

IFAD EC Drylands Restoration Team Meeting
6-8 May, Nairobi, Kenya

Introduction

A dataset, in various form, is a by-product of many research activities. Unfortunately, data curation is often neglected in comparison to the writing of an article or a report.

During the research, it is common to organize data in spreadsheets in a way which makes them easily understandable for the author at that specific time.

However the presentation of the data is very important to improve the spreading of the research, even if the dataset is not the primary product.

If the dataset is going to be released to the public, we must follow machine-readable standards, assuring that any future user can read it, understand its content, and use the data in other research projects.

Preliminary Steps

“When you are working with spreadsheets, during data clean up or analyses, it’s very easy to end up with a spreadsheet that looks very different from the one you started with” (Bahlai, 2017).

In order to avoid errors and data losses, do not modify the original dataset. Create a new copy to curate.

Enhance the title of the new dataset, providing some context.

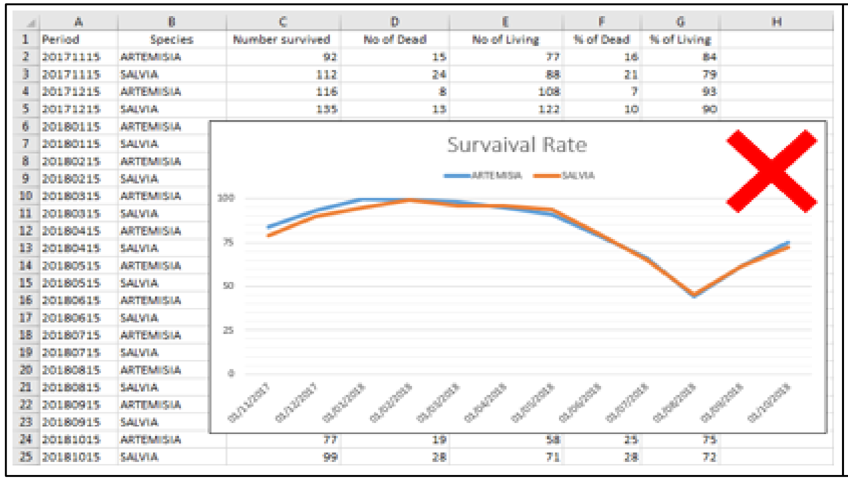
Old Dataset Title:	Rangeland Species Composition
Enhanced Dataset Title:	Annual and Perennial Rangeland Plant Cover and Species Composition, Tatatouine, Tunisia, November 2018

If the data is from a Primary Article Citation, use the naming convention "Data from: title of the article" (USDA, 2016).

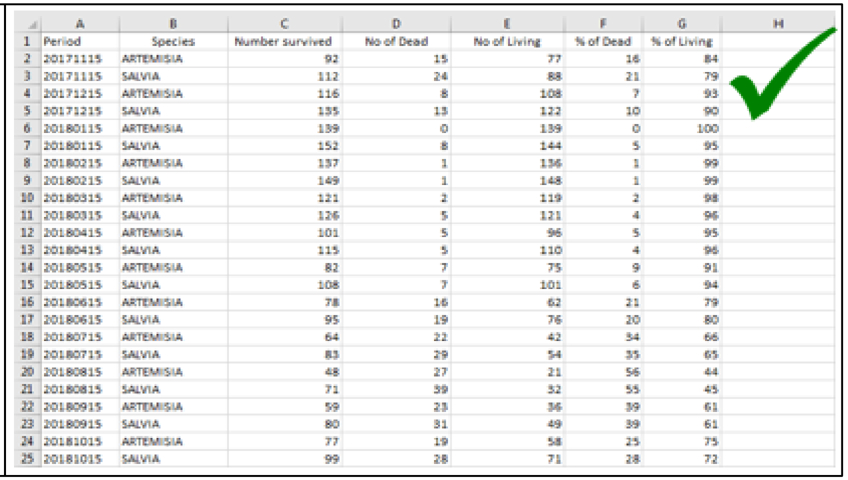
Data Curation – Elaboration Management

The dataset should contain only raw data. Each elaboration is subject to error. Providing only raw data, we enable future users to use them for their research, avoiding the risk to replicate elaboration errors.

- No graphs
- No formulas
- No percentages
- No 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	

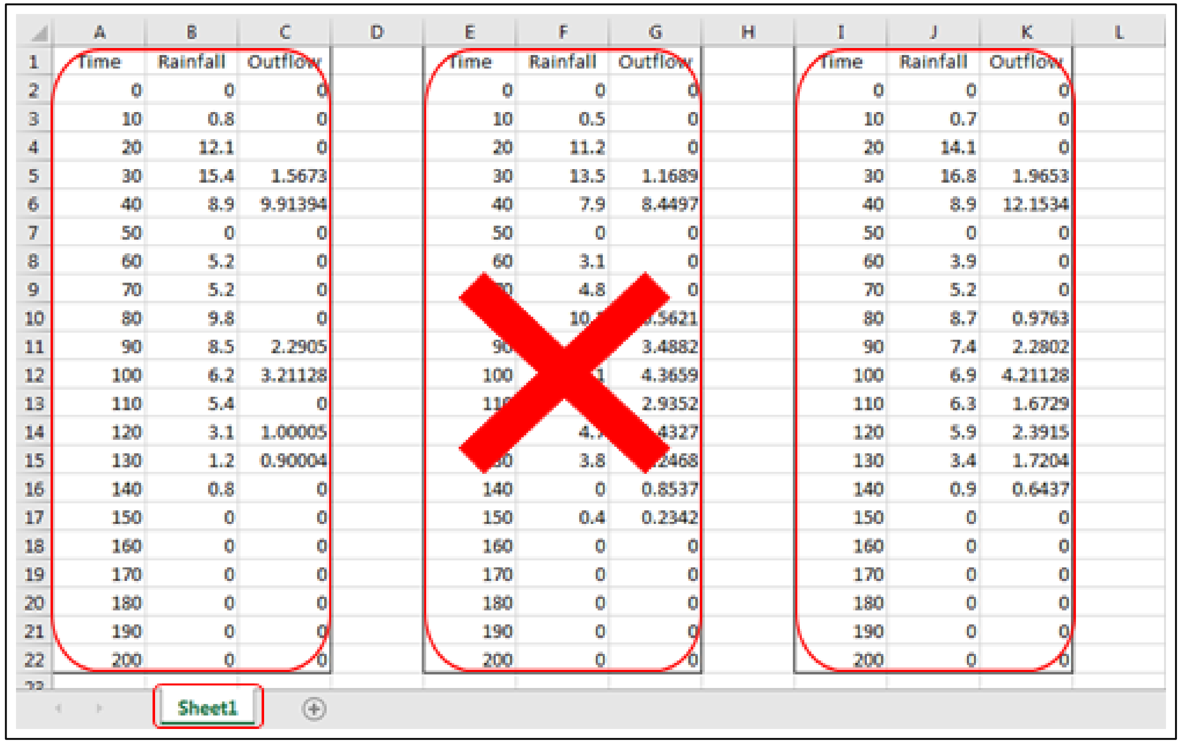


	A	B	C	D	E	F	G	H
1	Period	Species	Number survived	No of Dead	No of Living	% of Dead	% of Living	
2	20171115	ARTEMISIA	92	15	77	16	84	
3	20171115	SALVIA	112	24	88	21	79	
4	20171215	ARTEMISIA	116	8	108	7	93	
5	20171215	SALVIA	135	13	122	10	90	
6	20180115	ARTEMISIA	139	0	139	0	100	
7	20180115	SALVIA	152	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	

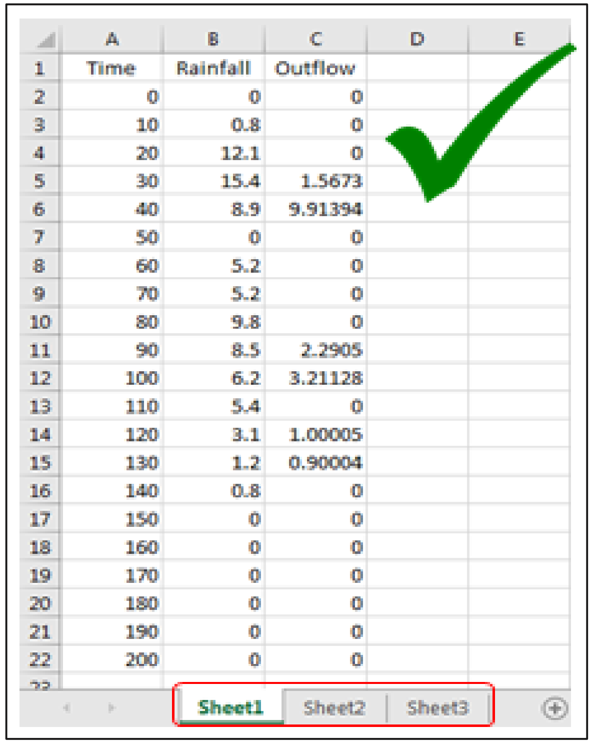
Data Curation – Tables Arrangement

It is not possible to have multiple data tables in one spreadsheet and use blank rows or columns to separate the data. It is not machine-readable.

Each spreadsheet must contain a single table.



	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		[illegible]	10.[illegible]	[illegible].5621		80	8.7	0.9763	
11	90	8.5	2.2905		90	[illegible]	3.4882		90	7.4	2.2802	
12	100	6.2	3.21128		100	[illegible]	4.3659		100	6.9	4.21128	
13	110	5.4	0		11[illegible]	[illegible]	2.9352		110	6.3	1.6729	
14	120	3.1	1.00005		[illegible]	4.[illegible]	[illegible]4327		120	5.9	2.3915	
15	130	1.2	0.90004		[illegible]0	3.8	[illegible]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	

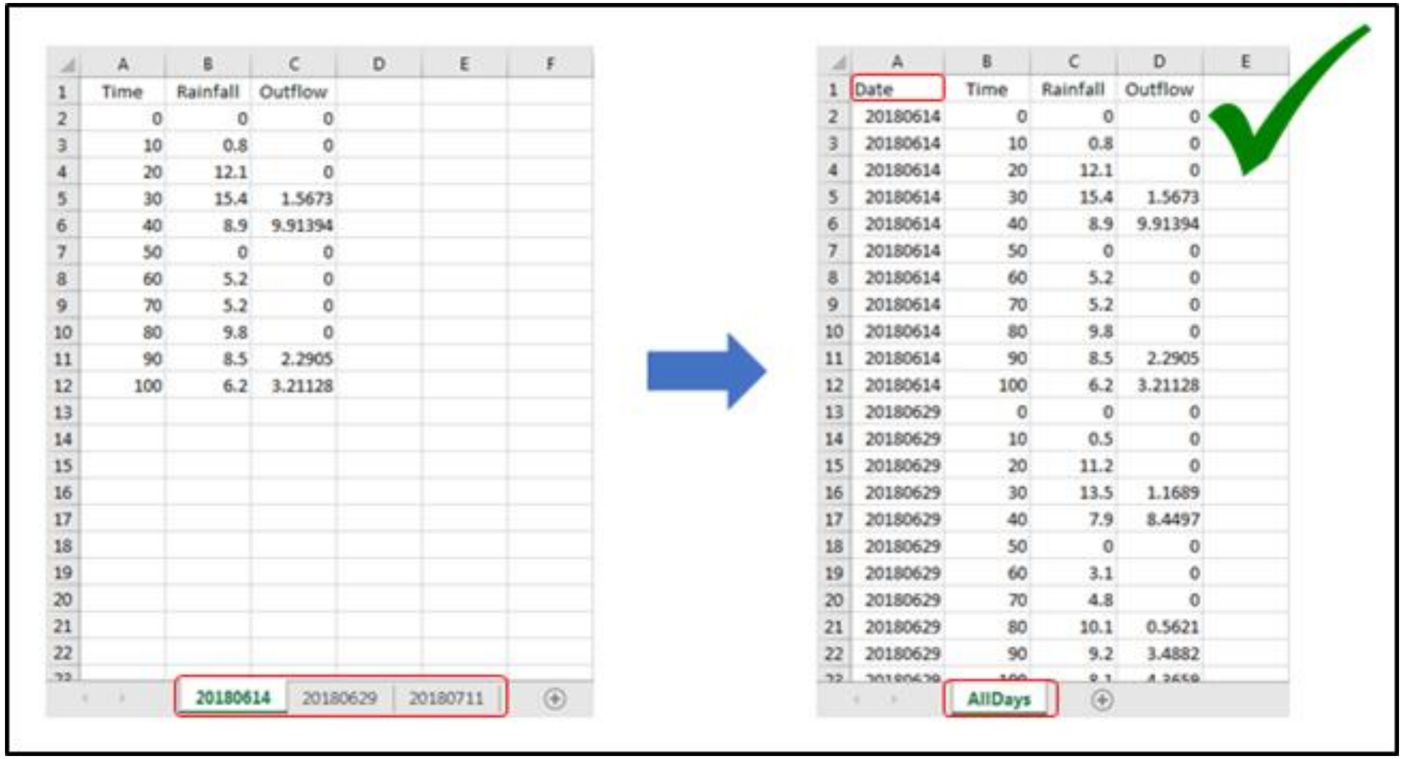


	A	B	C	D	E
1	Time	Rainfall	Outflow		
2	0	0	0		
3	10	0.8	0		
4	20	12.1	0		
5	30	15.4	1.5673		
6	40	8.9	9.91394		
7	50	0	0		
8	60	5.2	0		
9	70	5.2	0		
10	80	9.8	0		
11	90	8.5	2.2905		
12	100	6.2	3.21128		
13	110	5.4	0		
14	120	3.1	1.00005		
15	130	1.2	0.90004		
16	140	0.8	0		
17	150	0	0		
18	160	0	0		
19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

Data Curation – Tables Arrangement

If different spreadsheets contain similar data, that need to be analysed together, they can be merged, allowing the computer to see the connection.

However, be sure to add a column defining possible differences, like date or location.



	A	B	C	D	E	F
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			

	A	B	C	D	E
1	Date	Time	Rainfall	Outflow	
2	20180614	0	0	0	
3	20180614	10	0.8	0	
4	20180614	20	12.1	0	
5	20180614	30	15.4	1.5673	
6	20180614	40	8.9	9.91394	
7	20180614	50	0	0	
8	20180614	60	5.2	0	
9	20180614	70	5.2	0	
10	20180614	80	9.8	0	
11	20180614	90	8.5	2.2905	
12	20180614	100	6.2	3.21128	
13	20180629	0	0	0	
14	20180629	10	0.5	0	
15	20180629	20	11.2	0	
16	20180629	30	13.5	1.1689	
17	20180629	40	7.9	8.4497	
18	20180629	50	0	0	
19	20180629	60	3.1	0	
20	20180629	70	4.8	0	
21	20180629	80	10.1	0.5621	
22	20180629	90	9.2	3.4882	
23	20180629	100	8.1	4.2659	

Data Curation – Data Management

The dataset structure should be clean and simple:

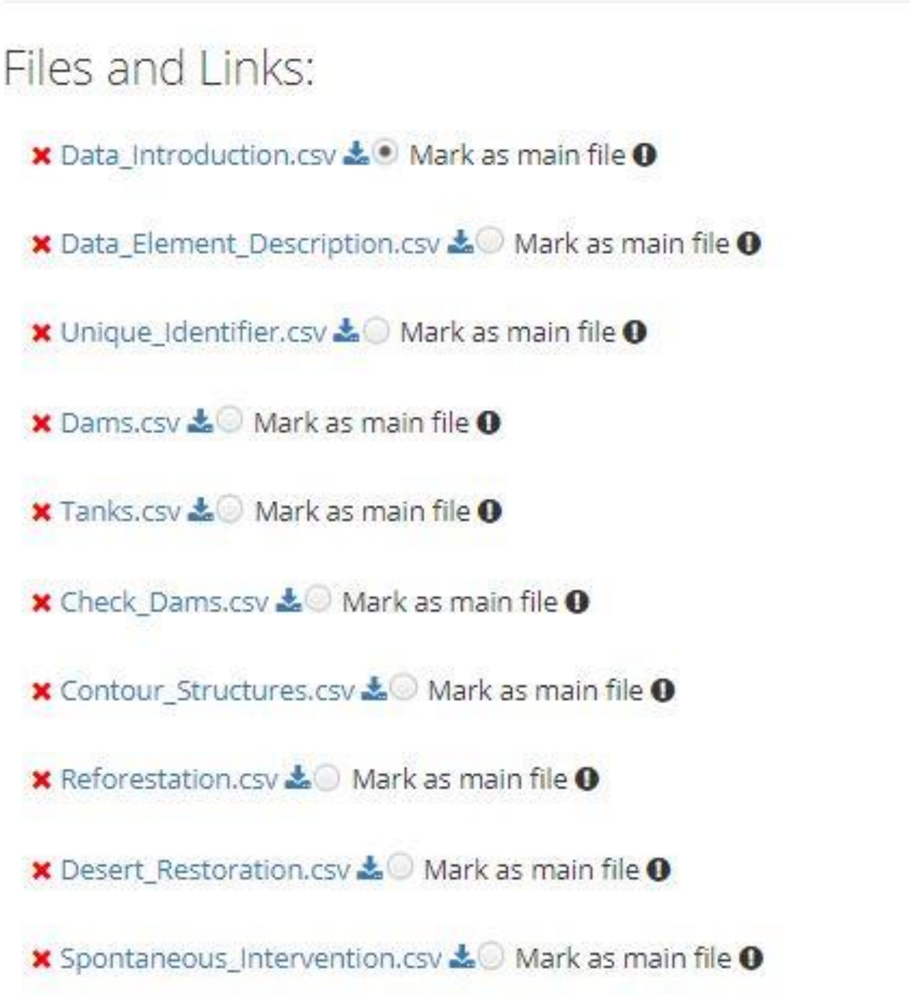
- Use short title for the head columns, without spaces or symbols. Just write in camel case or use underscore
- Only 1 information for each cell
- No vague or misleading information, when possible use numeric code
- Use ISO standards for date and time (YYYYMMDDHHMM)
- No merged cells
- No comments, use a column for the notes
- No empty cells. Use NA for missing or null data
- Although text format and colour could improve the readability they must be avoided, because they can easily be lost during transfer, compromizing the overall structure

File Format

Whatever software we are using we must be sure the files will be readable in the future, using different version of a licensed or unlicensed software.

The CSV or comma separated value files are the preferred data format for most of data repositories and are the recommended one for publishing machine-readable tabular data.

A CSV file contains a single spreadsheet: a dataset uploaded in MEL will be a collection of several CSV files.



Files and Links:

- ✖ Data_Introduction.csv ● Mark as main file
- ✖ Data_Element_Description.csv ○ Mark as main file
- ✖ Unique_Identifier.csv ○ Mark as main file
- ✖ Dams.csv ○ Mark as main file
- ✖ Tanks.csv ○ Mark as main file
- ✖ Check_Dams.csv ○ Mark as main file
- ✖ Contour_Structures.csv ○ Mark as main file
- ✖ Reforestation.csv ○ Mark as main file
- ✖ Desert_Restoration.csv ○ Mark as main file
- ✖ Spontaneous_Intervention.csv ○ Mark as main file

Data Dictionary – Dataset Introduction

The Dataset Introduction provide an overall explanation about the dataset scope and creation. It must include:

- Description: 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)
- Summary: A shorter description of the dataset, usually no more than a sentence or two (USDA, 2016)
- Start_Date: The date in which the data collection starts
- End_Date: The date in which the data collection ends
- Author: Dataset first author
- CoAuthor: Dataset co-authors

Data Dictionary – Dataset Introduction

Description	Summary	Start_Date	End_Date	Author	CoAuthor
A rich free text description that provides as much explanation as possible about the dataset.	A shorter description of the dataset, usually no more than a sentence or two.	YYYYMMDD	YYYYMMDD	Dataset first author	Dataset co-authors

The tab structure is customizable according to the needs of the author. It can include specific section about geographical location, metodologies and additional notes.

Data Dictionary – Element Description

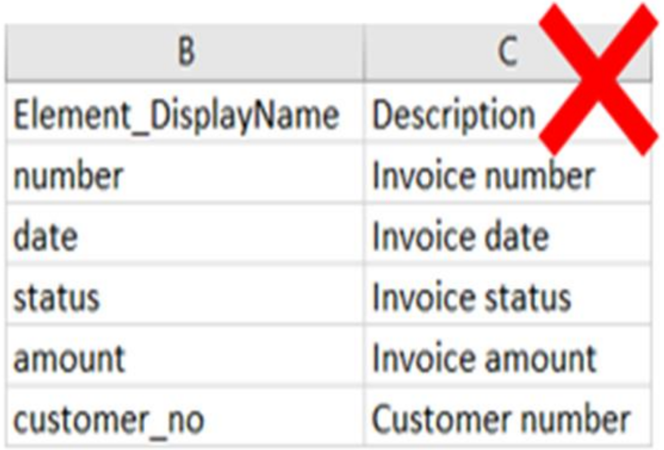
This is the most important component of the Data Dictionary. It provides explanation about the meaning of each variable and correspondences for any code used.

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

Data Dictionary – Element Description

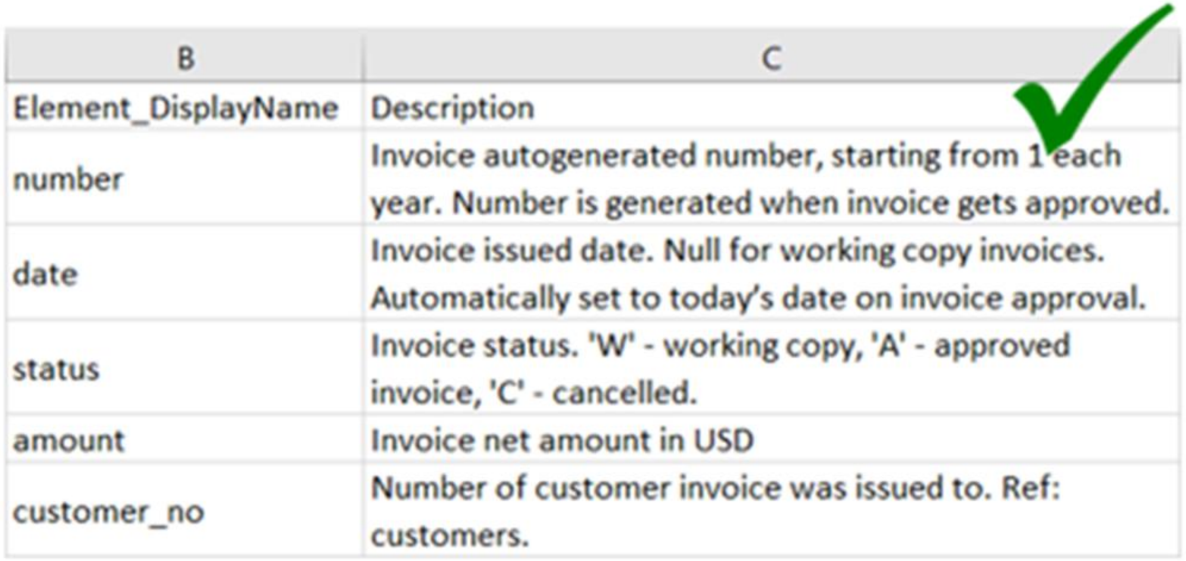
The suggested template for structuring manually the “Dataset Elements Description” includes the following fields (USDA, 2016):

- Spreadsheet_Tab: The tab where is the element
- Element_DisplayName: The dataset element name
- Description: A brief and complete element definition that could stand alone from other elements definition



✗

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

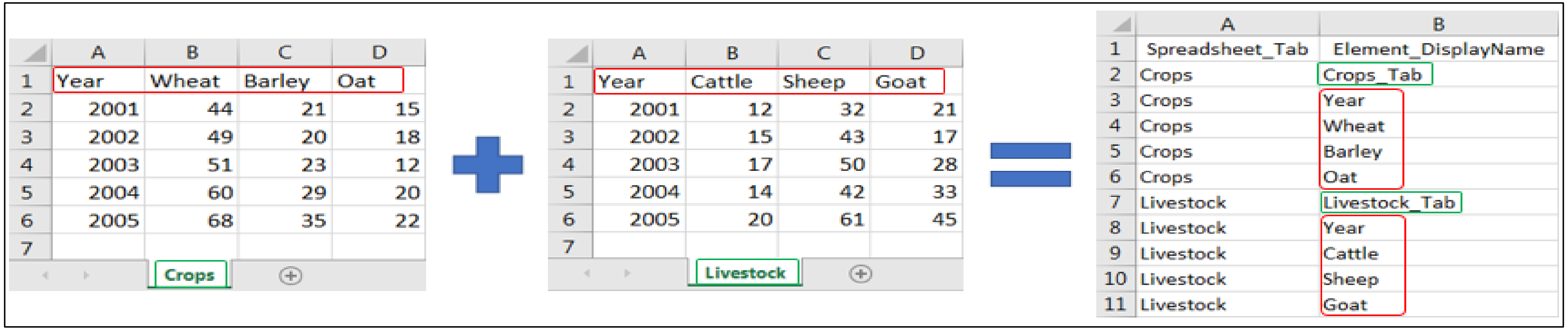


✓

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.

Data Dictionary – Element Description

- Unit: The unit of measurement adopted for the elements
- Data_Type: The type of data values contained in the field (e.g. varchar, integer, date, etc.)
- Character_Length: The length of data values contained in the field (maximum length for Excel is 255)
- Acceptable_Values: The list of acceptable values in this field. In some case it can be also a range of values
- Required: Express the requirement of values in the field for dataset status and validity
- Accepts_NullValue: Express the possibility of null values in the corresponding dataset field



	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

If the dataset includes multiple tabs, the various element of each tab should be listed in the same Data Dictionary file.

Data Dictionary – Unique Identifier

We often assume the use of certain terms to be clear, but it is not always the case, especially outside our research team.

To make sure to solve any possible ambiguity, in the unique identifier tab are reported the corresponding link for the dataset terms and concepts to the on-line thesaurus. This is very useful to avoid any misunderstanding on the elements (plant species, animals, etc.) analyzed and reported in the set of data (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

Final Recommendation

http://repo.mel.cgiar.org/handle/20.500.11766/9400

The General Dataset Curation Guide is available as a final draft at the link above. Each one of you can read it and start applying it to his own work until it will become common practice.

The Guide defines a standard, however it is still a work in progress, and certain types of file may need more curation work than others or even *ad hoc* solution. You can signal potential issues to the data curation staff, helping to expand the future version of the guide.

Thanks for your attention!