RESEARCH PROGRAM ON Grain Legumes and Dryland Cereals

GEOAGRO

ICARDA
Science for resilient livelihoods in dry areas

CGIAR Platform for Big Data in Agriculture

General Dataset Curation Guide

Francesco Bonechi

RESEARCH PROGRAM ON Livestock

MEL monitoring evaluation and learning

Introduction

- “Many valuable datasets are poorly curated, which contributes to errors, redundant effort, and obstacles to replication and use” (Ruggles, 2018)

- It is common to organize data in spreadsheets in a way which makes them easily understandable for the dataset author at that time, without following the machine-readable standards or considering any next research use.

Data Curation Role

“Data curation activities enable data discovery and retrieval, maintain data quality, add value, and provide for re-use over time” (Munoz, 2017)

Nowadays specific jobs related to data curation responsibilities are increasing (with title like “data curator” or “data curation specialist”)

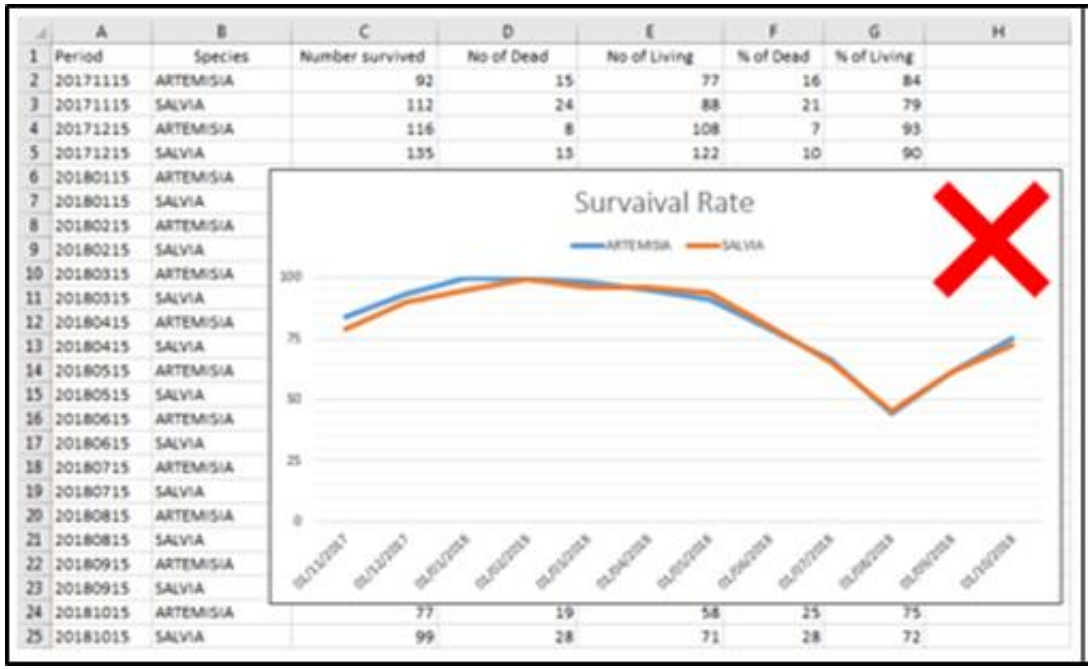
Preliminary Steps

1. In order to be able to reproduce your analyses and be sure you don't lose any data within curation processes, don't modify the original dataset but first create a new copy of your original data where to work on.
2. Enhance the dataset title of the new file "entering a descriptive title including dates, locations, and specific metrics that make the dataset unique" (USDA, 2016).

Old Dataset Title:	Rangeland Species Composition
Enhanced Dataset Title:	Annual_and_Perennial_Rangeland_PlantCover_and_SpeciesComposition_Pavlodar_Kazakhstan_November2018

Data Calculations and Summaries

1. Graphics and figures must be removed from the dataset spreadsheet tab.

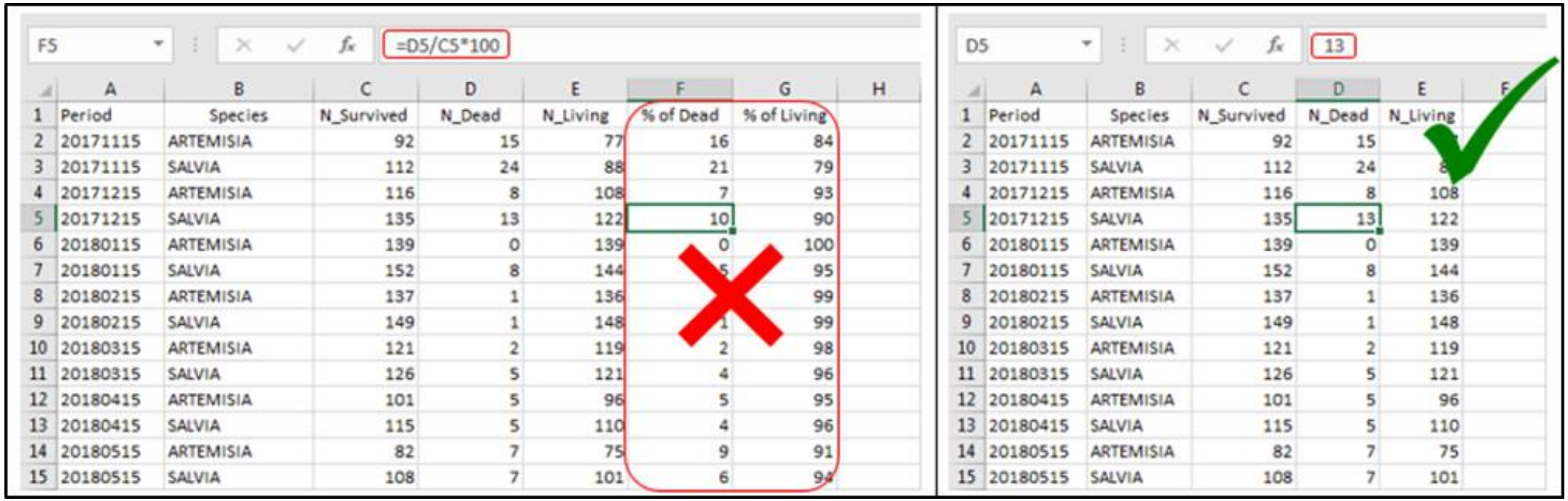


	A	B	C	D	E	F	G
1	Period	Species	Number survived	No of Dead	No of Living	% of Dead	% of Living
2	20171115	ARTEMISIA	92	15	77	16	84
3	20171115	SALVIA	112	24	88	21	79
4	20171215	ARTEMISIA	116	8	108	7	93
5	20171215	SALVIA	135	13	122	10	90
6	20180115	ARTEMISIA					
7	20180115	SALVIA					
8	20180215	ARTEMISIA					
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
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 Calculations and Summaries

2. Formulas and any other type of elaborations must be removed from the dataset spreadsheet tab. This may force to delete entire spreadsheet columns.



F5 =D5/C5*100

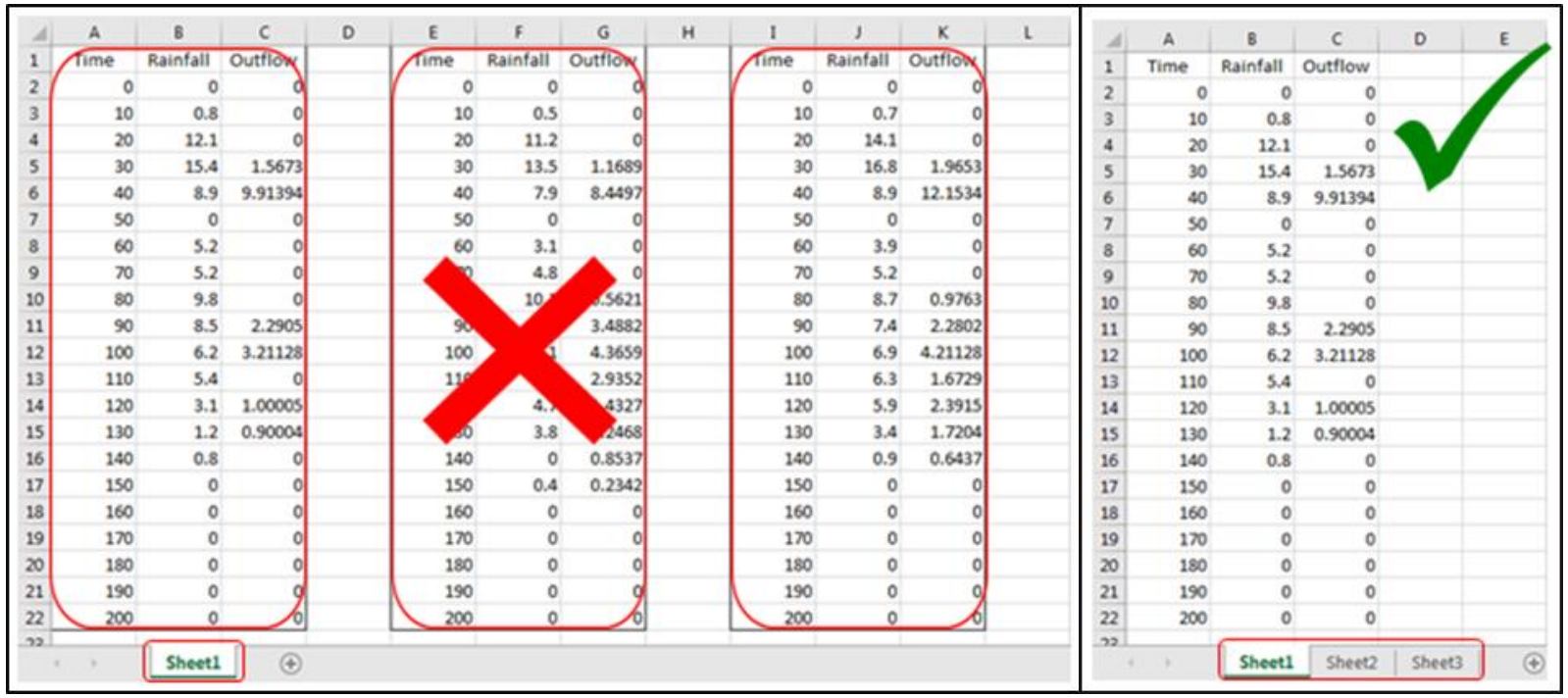
	A	B	C	D	E	F	G	H
1	Period	Species	N_Survived	N_Dead	N_Living	% of Dead	% of Living	
2	20171115	ARTEMISIA	92	15	77	16	84	
3	20171115	SALVIA	112	24	88	21	79	
4	20171215	ARTEMISIA	116	8	108	7	93	
5	20171215	SALVIA	135	13	122	10	90	
6	20180115	ARTEMISIA	139	0	139	0	100	
7	20180115	SALVIA	152	8	144	5	95	
8	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	

D5 13

	A	B	C	D	E	F
1	Period	Species	N_Survived	N_Dead	N_Living	
2	20171115	ARTEMISIA	92	15		
3	20171115	SALVIA	112	24	8	
4	20171215	ARTEMISIA	116	8	108	
5	20171215	SALVIA	135	13	122	
6	20180115	ARTEMISIA	139	0	139	
7	20180115	SALVIA	152	8	144	
8	20180215	ARTEMISIA	137	1	136	
9	20180215	SALVIA	149	1	148	
10	20180315	ARTEMISIA	121	2	119	
11	20180315	SALVIA	126	5	121	
12	20180415	ARTEMISIA	101	5	96	
13	20180415	SALVIA	115	5	110	
14	20180515	ARTEMISIA	82	7	75	
15	20180515	SALVIA	108	7	101	

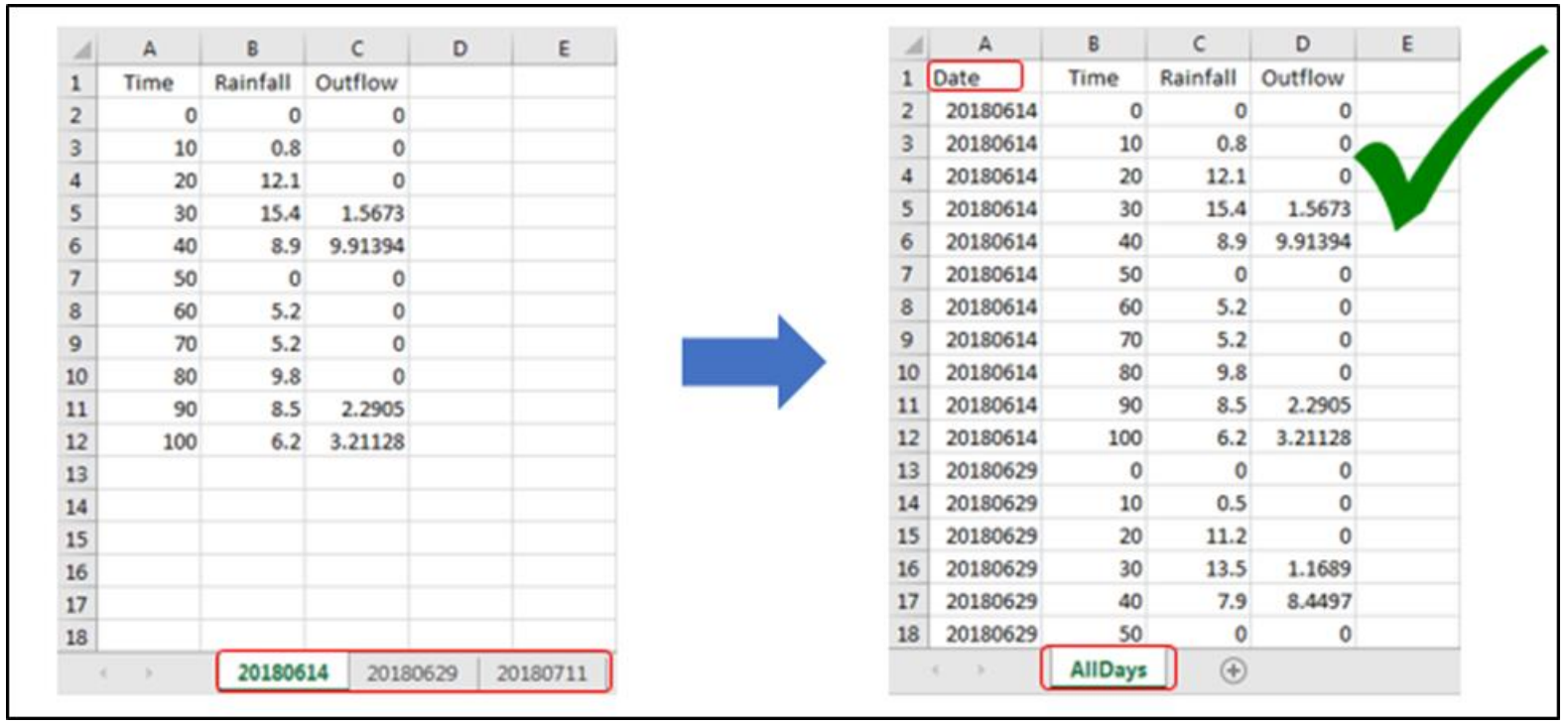
Data Table

1. In order to avoid false association during data system readability, blank rows and columns must not be used to separate the dataset in different tables or sections.



Data Table

Note: Improving the dataset column arrangements is possible to reduce the spreadsheet tab number.



	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					
14					
15					
16					
17					
18					

20180614 | 20180629 | 20180711

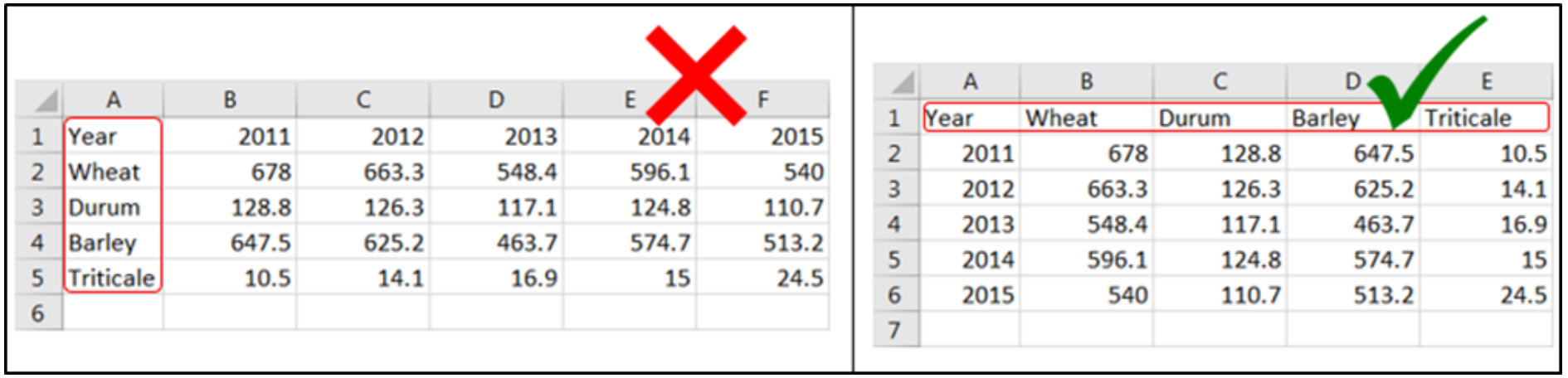
	A	B	C	D	E
1	Date	Time	Rainfall	Outflow	
2	20180614	0	0	0	
3	20180614	10	0.8	0	
4	20180614	20	12.1	0	
5	20180614	30	15.4	1.5673	
6	20180614	40	8.9	9.91394	
7	20180614	50	0	0	
8	20180614	60	5.2	0	
9	20180614	70	5.2	0	
10	20180614	80	9.8	0	
11	20180614	90	8.5	2.2905	
12	20180614	100	6.2	3.21128	
13	20180629	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	

AllDays

Structuring Data in the Spreadsheet

"The cardinal rules of using spreadsheet programs for data are" (Bahlai, 2017):

- Put all the variables in columns. Each column corresponds to a variable.
- Put each observation in its own row. Each row corresponds to an observation.
- Don't combine multiple pieces of information in one cell. Each cell corresponds to one value (data).



	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						

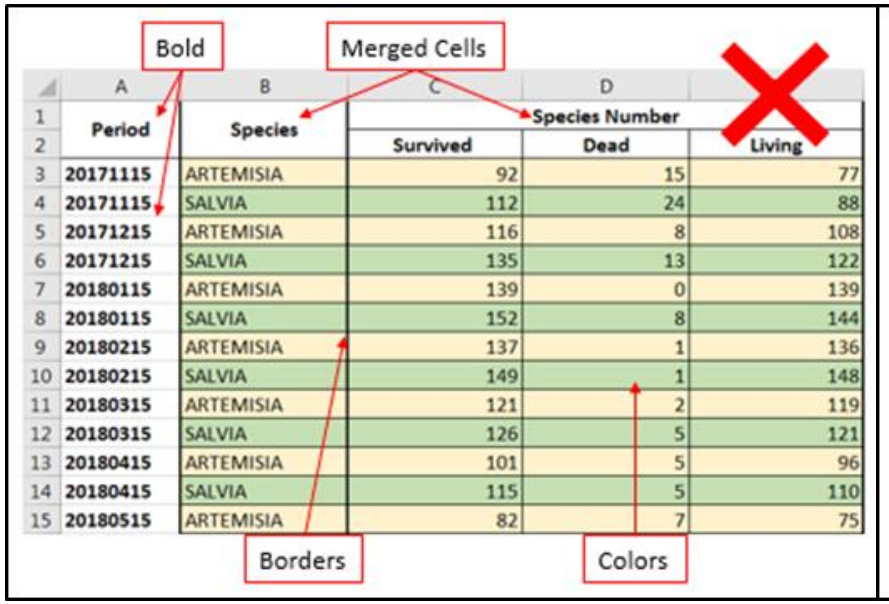
	A	B	C	D	E
1	Year	Wheat	Durum	Barley	Triticale
2	2011	678	128.8	647.5	10.5
3	2012	663.3	126.3	625.2	14.1
4	2013	548.4	117.1	463.7	16.9
5	2014	596.1	124.8	574.7	15
6	2015	540	110.7	513.2	24.5
7					

Structuring Data in the Spreadsheet

Formatting Features

The special formatting features (merged cells, boarders, colors, bold, etc.) must be avoid as much as possible.

“Consider restructuring your data in such a way that you will not need to merge cells or other esthetic features to organize your data” (Bahlai, 2017)



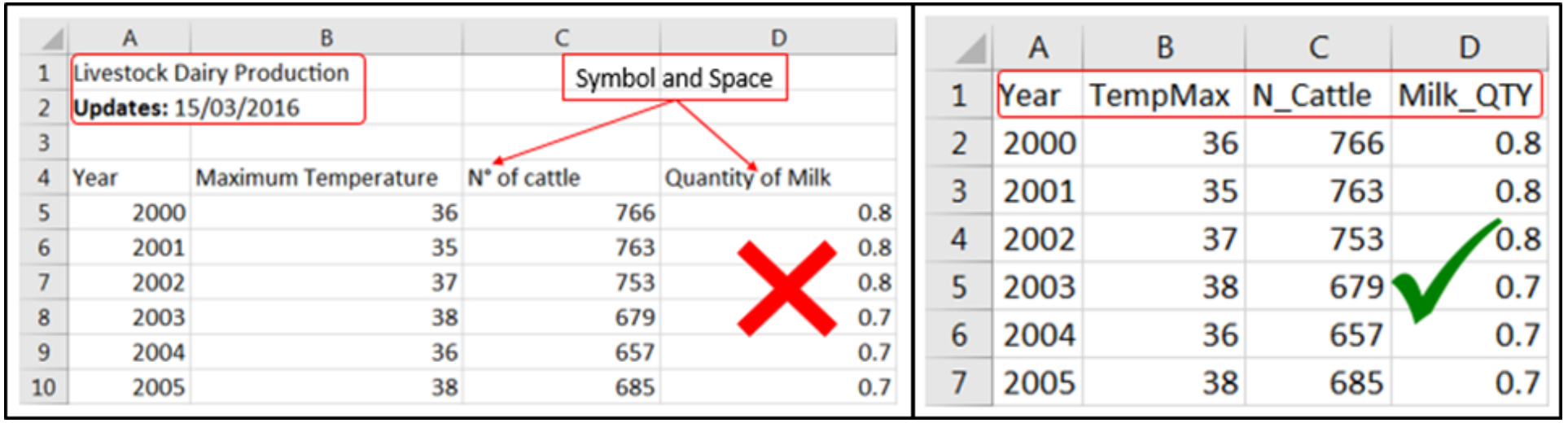
	A	B	C	D	E
1	Period	Species	N_Survived	N_Dead	N_Living
2	20171115	ARTEMISIA	92	15	77
3	20171115	SALVIA	112	24	88
4	20171215	ARTEMISIA	116	8	108
5	20171215	SALVIA	135	13	122
6	20180115	ARTEMISIA	139	0	139
7	20180115	SALVIA	152	8	144
8	20180215	ARTEMISIA	137	1	136
9	20180215	SALVIA	149	1	148
10	20180315	ARTEMISIA	121	2	119
11	20180315	SALVIA	126	5	121
12	20180415	ARTEMISIA	101	5	96
13	20180415	SALVIA	115	5	110
14	20180515	ARTEMISIA	82	7	75
15	20180515	SALVIA	108	7	101

Structuring Data in the Spreadsheet

Column Headers

Do not capture documentation and text descriptions in the tables themselves. The descriptive information can be recorded in the data dictionary or put into a “Note” column created for this purpose (USDA, 2017).

Ensure column headings do not contain spaces, hyphens or any other symbols. Only the underscore is allowed.



	A	B	C	D
1	Livestock Dairy Production			
2	**Updates:** 15/03/2016			
3				
4	Year	Maximum Temperature	N° of cattle	Quantity of Milk
5	2000	36	766	0.8
6	2001	35	763	0.8
7	2002	37	753	0.8
8	2003	38	679	0.7
9	2004	36	657	0.7
10	2005	38	685	0.7

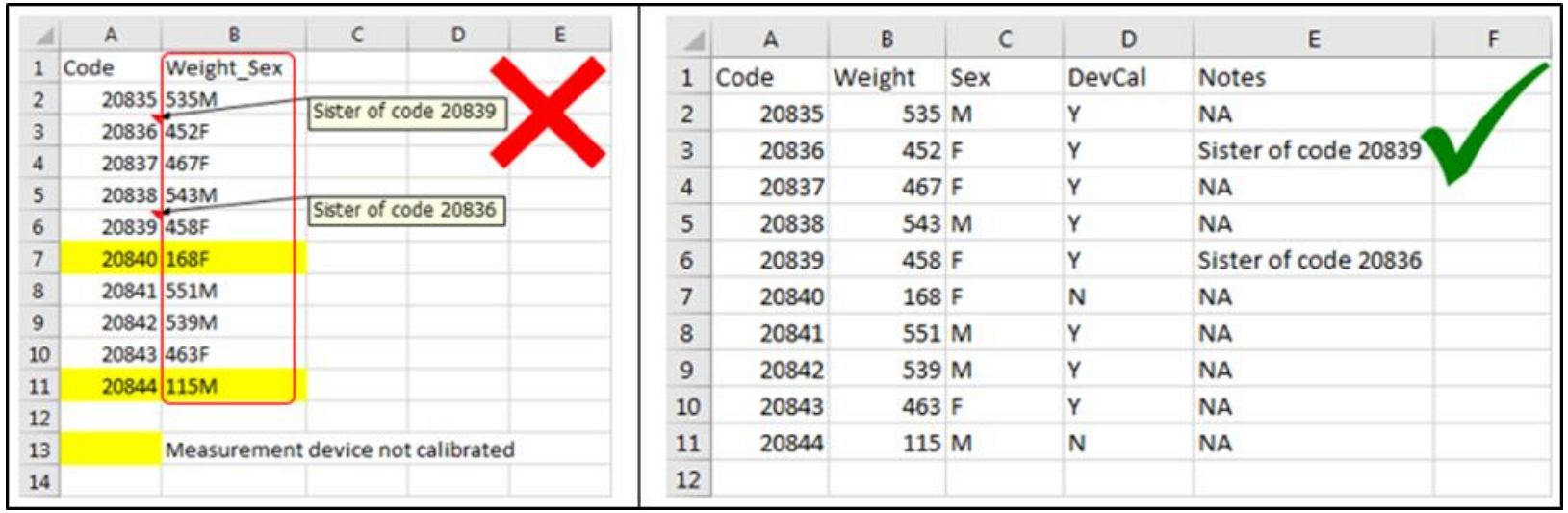
	A	B	C	D
1	Year	TempMax	N_Cattle	Milk_QTY
2	2000	36	766	0.8
3	2001	35	763	0.8
4	2002	37	753	0.8
5	2003	38	679	0.7
6	2004	36	657	0.7
7	2005	38	685	0.7

“Underscores (_) are a good alternative to spaces. Consider writing names in camel case (e.g. TestName) to improve readability” (Bahlai, 2017).

Structuring Data in the Spreadsheet

Data Entry

1. Data must be entered in a consistent way using always the same code for the same value.
2. No more than one piece of information can be in one cell.
3. Highlighted cells and comments must be removed since they may create problems for machine readability. These observations can be entered in new columns (e.g. Notes, etc.)



	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					

	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						

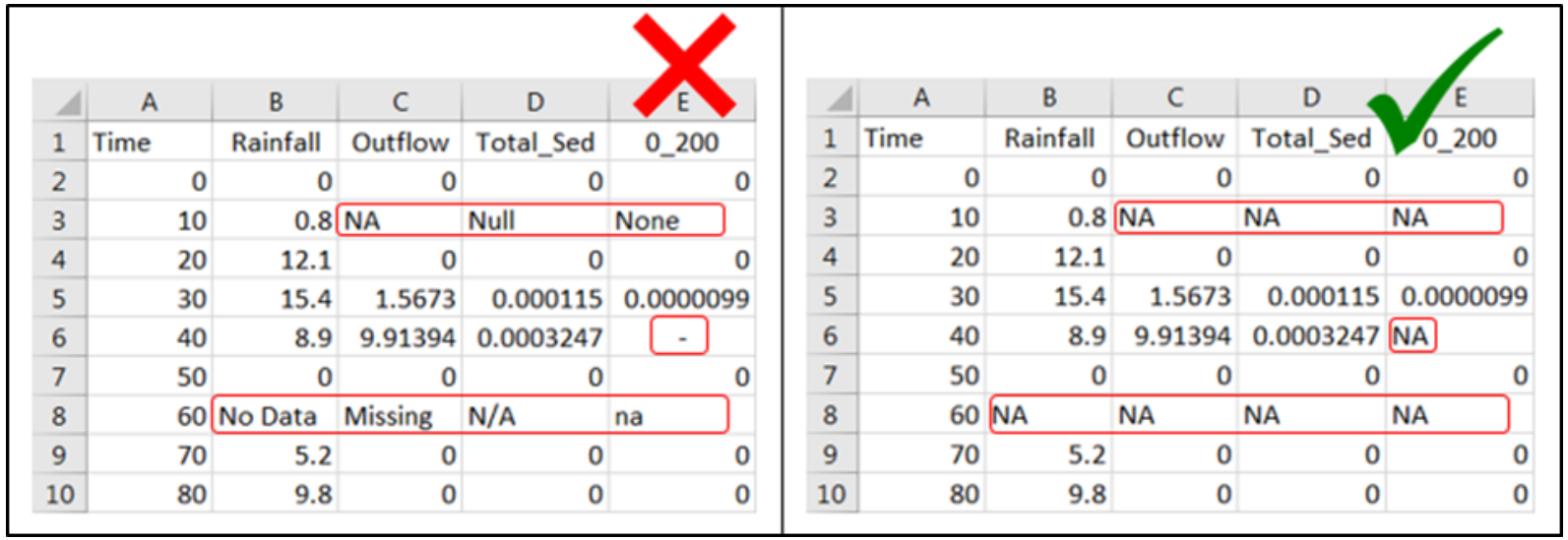
Note: When writing text in the cells, they can only contain text and spaces.

Structuring Data in the Spreadsheet

Null Values

Null values must be represented differently from “0”. In fact, “0” corresponds to a measured data while null value means that the data is not measured at all.

It is common to represent null values with blank cells but even to enter “NA” is a good option.



	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

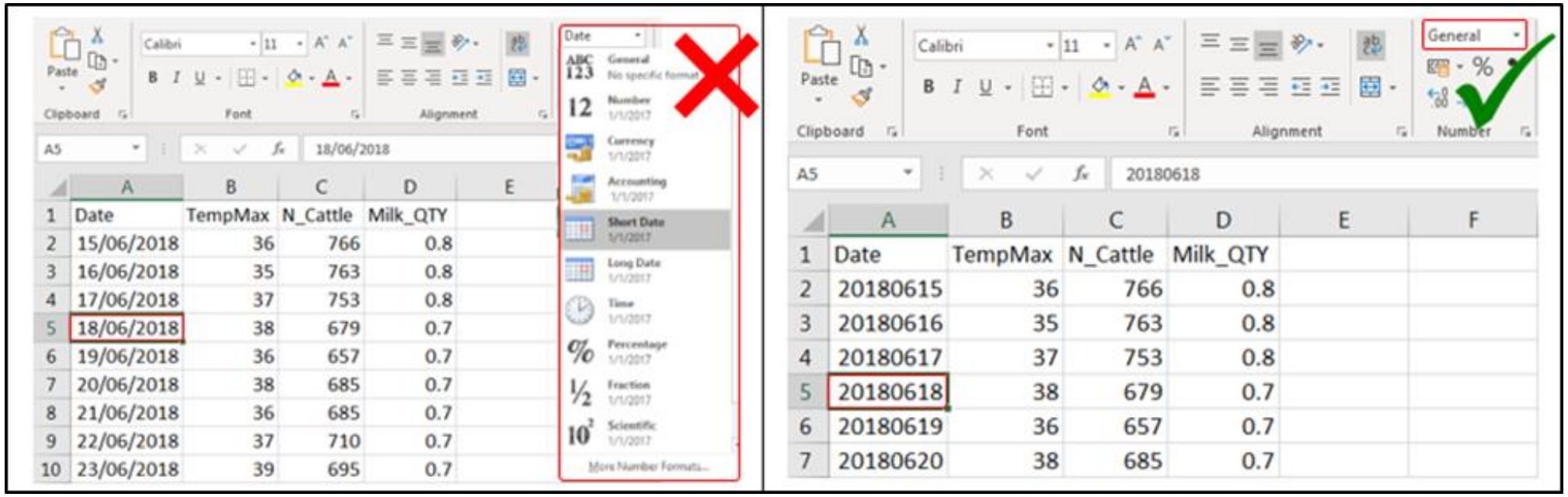
	A	B	C	D	E
1	Time	Rainfall	Outflow	Total_Sed	0_200
2	0	0	0	0	0
3	10	0.8	NA	NA	NA
4	20	12.1	0	0	0
5	30	15.4	1.5673	0.000115	0.0000099
6	40	8.9	9.91394	0.0003247	NA
7	50	0	0	0	0
8	60	NA	NA	NA	NA
9	70	5.2	0	0	0
10	80	9.8	0	0	0

Note: As long as the missing value representation is consistent and documented, the next users can replace the choice for a null value independently (Zwicker, 2016).

Structuring Data in the Spreadsheet

Dates and Time

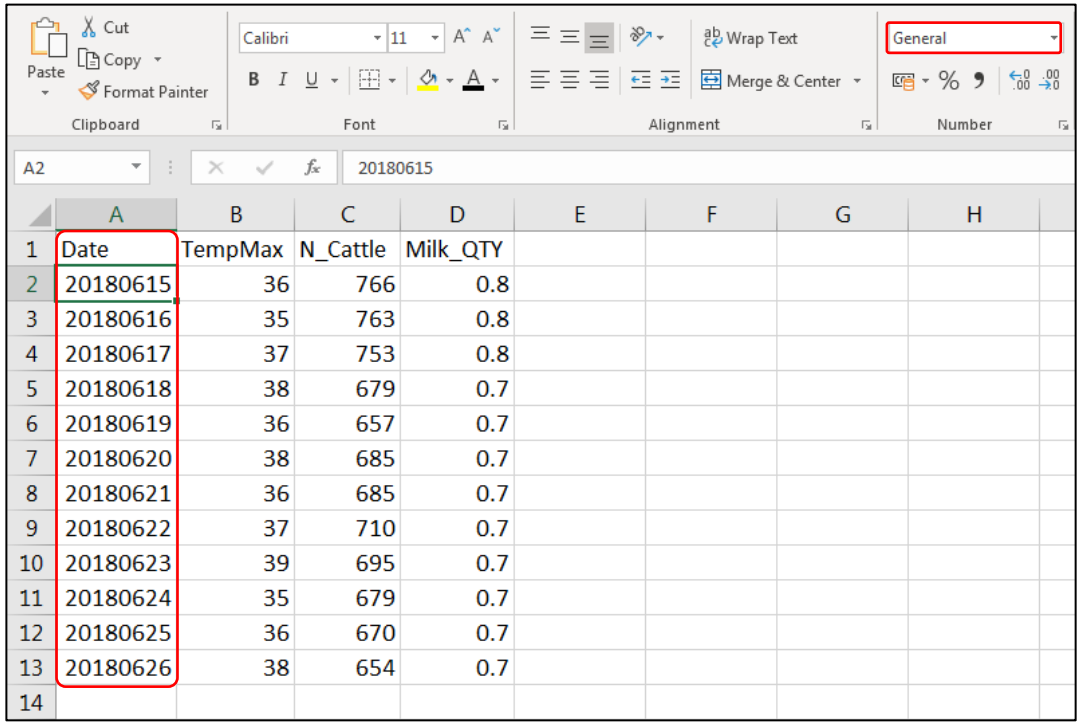
The special functionalities available must not be used since they are usually guaranteed to be compatible only within the same family of products (Bahlai, 2017).



Structuring Data in the Spreadsheet

Dates and Time

Basing on ISO standards (ISO 8601:2004), the suggested format to store dates is YYYYMMDD while for time is hhmmss using the 24-hours notation (which become YYYYMMDDhhmmss when represented together).



	A	B	C	D	E	F	G	H
1	Date	TempMax	N_Cattle	Milk_QTY				
2	20180615	36	766	0.8				
3	20180616	35	763	0.8				
4	20180617	37	753	0.8				
5	20180618	38	679	0.7				
6	20180619	36	657	0.7				
7	20180620	38	685	0.7				
8	20180621	36	685	0.7				
9	20180622	37	710	0.7				
10	20180623	39	695	0.7				
11	20180624	35	679	0.7				
12	20180625	36	670	0.7				
13	20180626	38	654	0.7				
14								

For example, June 15, 2018 17:25:35 become 20180615172535.

Structuring Data in the Spreadsheet

Latitudes and Longitudes

The recommended standard to report latitudes and longitudes data is using decimal degrees (DD), since they guarantee the possibility of treating latitude and longitude as a simple and numeric value facilitating any next software interpretations (Callahan, 2009).

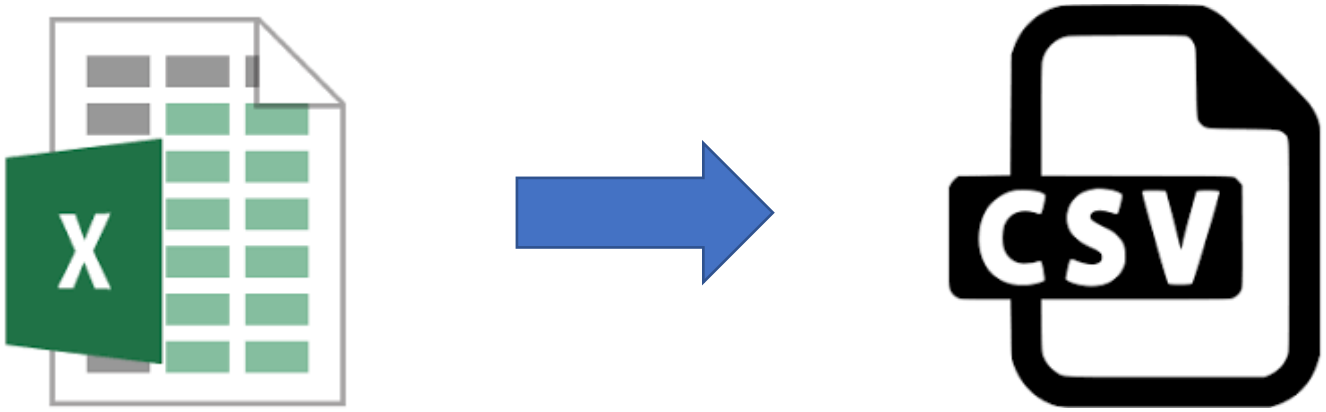
Unit	Latitude	Longitude
Decimal Degrees (DD)	**40.75889**	**-73.98513**
Degrees Minutes and Seconds (DMS)	40° 45' 32.004" N	73° 59' 6.468" W
Degrees Decimal Minutes (DM)	40° 45.5334'	-73° 59.1078'

	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

Stable File Format

From Excel to CSV

It is important to save the dataset in a consistent format that can be read well into the future and is independent of changes in applications. The CSV or comma separated value format is the preferred format for most of data repositories and are the recommended one for publishing machine-readable tabular data.



Data Dictionary

Data dictionaries are used to provide detailed information about the contents of a dataset or database, such as the names of measured variables, their data types or formats, and text descriptions. In particular (Briny, 2015):

- It helps the dataset author to remember all the details about the data along the years;
- It facilitates the dataset sharing with collaborators helping them to understand and use the data files.
- It helps for personnel who are “totally unfamiliar with the data, to pick up that data, understand and reproduce the results or reuse these for new research” (Briny, 2015) activities improving the credit of the dataset.

Data Dictionary

When the data are managed in professional databases, it is possible to automatically generate data dictionaries by the available tools in the software.

While, when data are managed in spreadsheets, text files, or comma separated values, the data dictionary must be created manually. A common approach when doing it manually, is to create three main files, to save in CSV format, which contains three different level of dataset information:

A. **Dataset Introduction:** where introductory and background explanatory information are reported;

B. **Dataset Elements Descriptions:** where the datasets fields are listed with their related information;

C. **Unique Identifier:** where the dataset elements and concepts are identified by dereferenceable URIs to online multilingual thesaurus.

Dataset Introduction

The purpose of the Dataset Introduction tab is to explain the contents of the dataset. Here are available the general information to make clear all the dataset aspects for next uses.

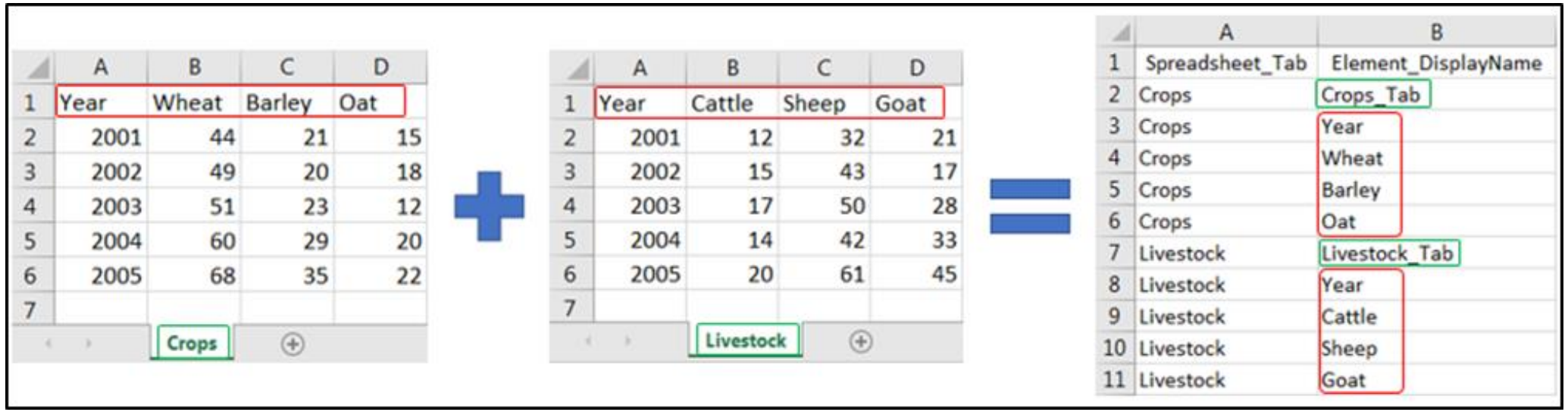
	A	B	C	D	E	F	G	H
1	Description	Summary	Start_Date	End_Date	Latitude	Longitude	Author	CoAuthor
2	A rich and full dataset description that explains how and why it was generated and informs how it should be used. Make sure that in this description are present the experiment settings (location, climatic conditions, etc.), data collection and processes methods, equipment used, 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 coordinate, in decimal degrees (DD), using WGS84 datum.	Longitude site coordinate, in decimal degrees (DD), using WGS84 datum.	Dataset first author.	Dataset co-author(s).

Dataset Elements Descriptions

This is the core of the data dictionary, since it is the document that allow the dataset users to fully understand the contents of the dataset (ORNL DAAC, 2018).

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

1. **Spreadsheet_Tab:** The tab where is the element.
2. **Element_DisplayName:** The dataset element name.



	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

Dataset Elements Descriptions

Description: A brief and complete element definition that could stand alone from other elements definitions.

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.

Source: Kononow, 2017

Note: It is important that descriptions are meaningful avoiding the text holding zero information.

Dataset Elements Descriptions

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
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

	A	B	C	D	E	F	G	H	I
1	Spreadsheet_Tab	Element_DisplayName	Description	Units	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. Retrieved from: 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	[2000,2015]	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_StudentsEducation	INS_StudentsEducation_tab	Data about male and female basic and secondary education in Tunisia. Last update: 17/03/2016. Source: Ministry of education (Statistique Tunisia). Retrieved from: http://www.ins.tn/en/themes/education.	NA	NA	NA	NA	NA	NA
7	INS_StudentsEducation	Years	The year to which this analysis refers.	yyyy	date	4	[1990,2014]	y	n
8	INS_StudentsEducation	BasicAndSecondaryEdu_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	[45000,55000]	n	y
9	INS_StudentsEducation	BasicAndSecondaryEdu_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	[45000,60000]	n	y

Source: Khawam, 2017a

Unique Identifier

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 online multilingual thesaurus.

	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	

Source: Khawam, 2017b

Conclusions

Dataset curation practices can be challenging depending on the status of the data. In general, they are based on a standardization of the dataset contents and the creation of the necessary documentation.

Following the practices mentioned before, starting from a dataset file in Excel format, we end with multiple files in CSV format. Thus, the result of this work is a .zip file containing the data dictionaries files and the dataset ones in CSV format.

Recommendations

All these aspects need to be considered from the first data collection steps; so that these practices become an integral part of the author's work thus reducing the heaviness of this activity as well as ensuring better results.

References

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/

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

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

Khawam, H., & Najjar, D. (2017b). *Statistics on Crops, with a Focus on Barley Production and Trade, in Tunisia*. Retrieved from https://hdl.handle.net/20.500.11766.1/YSTDVL

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). *Frequently asked questions about data curation*. Retrieved December 5, 2018, from https://guide.dhcuration.org/faq/

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

References

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

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

Thanks for your attention!