



Importance and principles of dataset curation

Pietro Bartolini, Asma Jeitani

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

icarda.org

International Center for Agricultural Research in the Dry Areas





Data curation: Why?

Data are the backbones of any scientific research.

Almost all research activities produce some kind of dataset as by-product.

Unfortunately a dataset is often considered less important than a paper or a report: We use it, organizing it according to our temporary needs, and then we leave it in some repository without any further preparation.

Without proper curation, a dataset risks to be completely un-usable for future research projects and it is impossible to easily spread the research results.

A curated dataset is a re-usable dataset that can be released to the public as an independent product.

Data Curation: How?

Curated dataset main characteristics:

- It provides relevant information about the content and the context of the research
- It contains only raw data, without additional elaboration
- It follows machine-readable standards

A detailed description of the curation process is available in *The General Dataset Curation Guide* in the link below:

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

Preliminary Steps

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

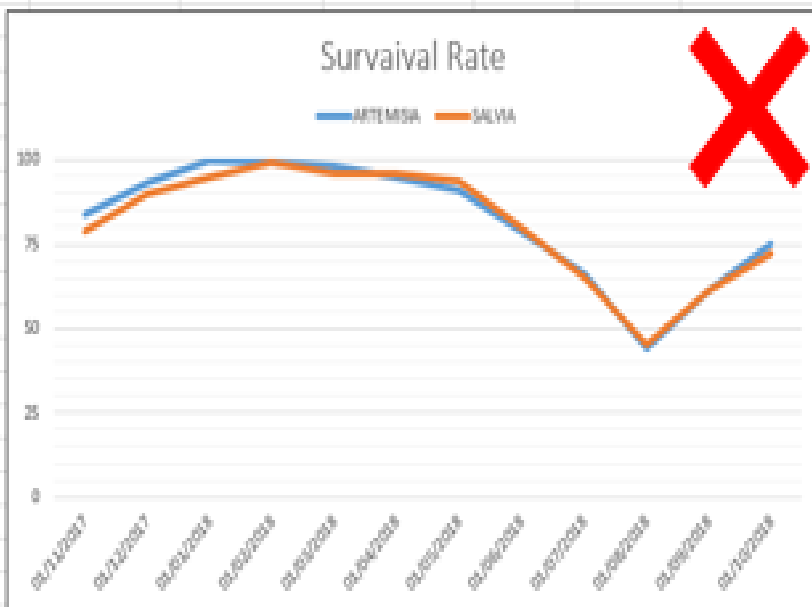
The final dataset should contain only raw data, since each elaboration is subject to error and other researchers may be interested in different analysis.

- No graphs
- No formulas
- No percentages
- No elaborations

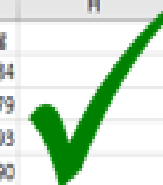
Providing only raw data, we enable future users to use them for their research, avoiding the risk to replicate elaboration errors.

Data Curation – Elaboration Management

	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						
25	20181015	SALVIA						



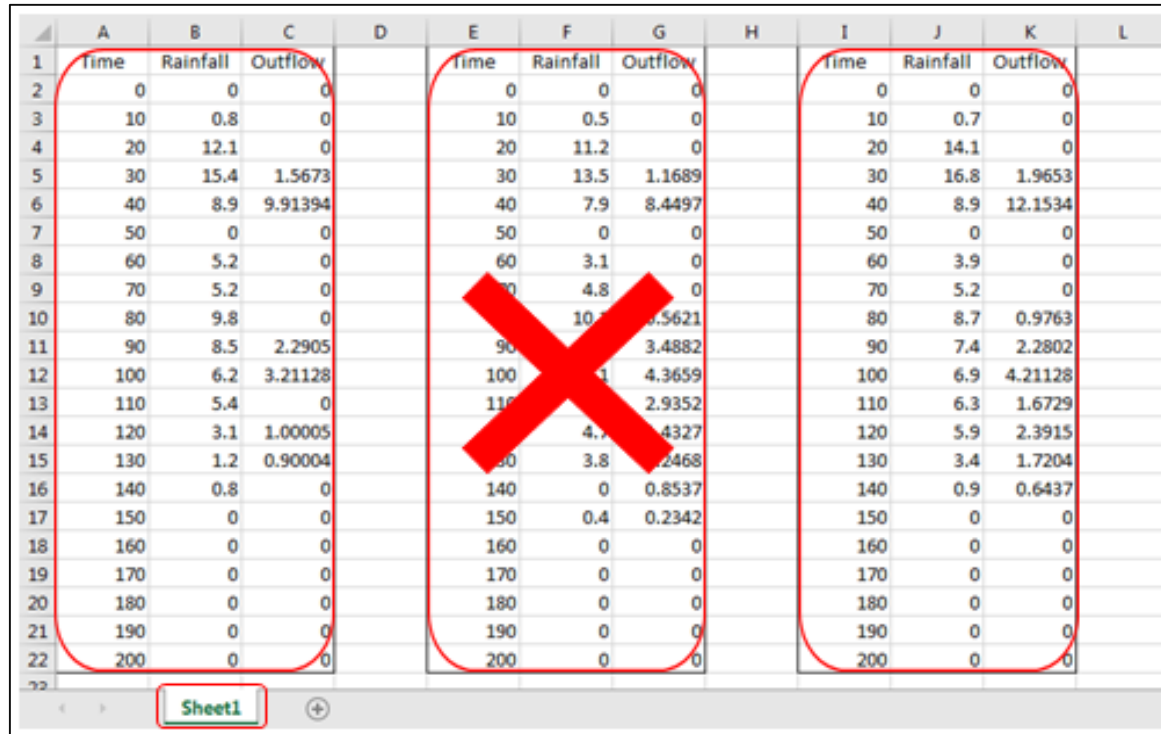
	A	B	C	D	E	F	G	H
1	Period	Species	Number survived	No of Dead	No of Living	% of Dead	% of Living	
2	20171115	ARTEMISIA	92	15	77	16	84	
3	20171115	SALVIA	112	24	88	21	79	
4	20171215	ARTEMISIA	116	8	108	7	93	
5	20171215	SALVIA	135	13	122	10	90	
6	20180115	ARTEMISIA	139	0	139	0	100	
7	20180115	SALVIA	152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

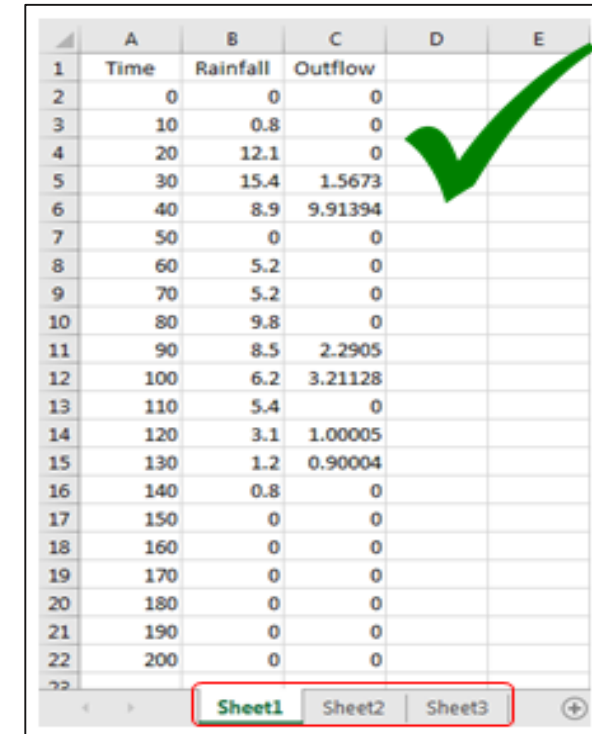
In order to be machine-readable, the dataset cannot have multiple data tables in a single spreadsheet, using blank rows or columns to separate the data.

Each spreadsheet must contain a single table.



The image shows a spreadsheet with three data tables arranged side-by-side in columns A-C, E-G, and I-K. Each table has headers 'Time', 'Rainfall', and 'Outflow'. The data is repeated for time intervals from 0 to 200. A large red 'X' is overlaid on the middle table, indicating this arrangement is incorrect for data curation.

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
110	5.4	0
120	3.1	1.00005
130	1.2	0.90004
140	0.8	0
150	0	0
160	0	0
170	0	0
180	0	0
190	0	0
200	0	0




The image shows a spreadsheet with a single data table in columns A-C. The table has headers 'Time', 'Rainfall', and 'Outflow'. The data is repeated for time intervals from 0 to 200. A large green checkmark is overlaid on the right side of the table, indicating this arrangement is correct for data curation.

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
110	5.4	0
120	3.1	1.00005
130	1.2	0.90004
140	0.8	0
150	0	0
160	0	0
170	0	0
180	0	0
190	0	0
200	0	0

Data Curation – Tables Arrangement

If different spreadsheets contain similar data and share a similar structure, they may need to be analysed together. In order to allow the program to see the connection, they can be merged in a single spreadsheet.



	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			
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						

	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	9.1	4.2658	

Note:
While merging the spreadsheets, be sure to add a column with a clear ID code, such as different dates or locations.



Data Curation – Data Management

The dataset structure should be clean and simple:

- Short title for the head columns (no spaces or symbols, 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
- No strange text format or colour (they could improve the readability, highlighting relevant data, but they can easily be lost during transfer, compromising the overall structure)

File Format

Whatever software we are using, we must be sure the files will be readable in the future, using different versions 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 ones for publishing machine-readable tabular data.

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

Files and Links:

- ✗ [DataDictionary_Introduction.csv](#)  ☒ Mark as main file 
- ✗ [DataDictionary_ElementDescription.csv](#)  ☐ Mark as main file 
- ✗ [DataDictionary_UniqueIdentifier.csv](#)  ☐ Mark as main file 
- ✗ [Bonga_Data_BWT.csv](#)  ☐ Mark as main file 
- ✗ [Bonga_Data_SMWT.csv](#)  ☐ Mark as main file 
- ✗ [Horro_Data_BWT.csv](#)  ☐ Mark as main file 
- ✗ [Horro_Data_SMWT.csv](#)  ☐ Mark as main file 
- ✗ [Menz_Data_BWT.csv](#)  ☐ Mark as main file 
- ✗ [Menz_Data_SMWT.csv](#)  ☐ Mark as main file 

Providing Context – Data Dictionary

In order to provide information about the content and the context of the research, each dataset must include a complete Data Dictionary:

- Data Dictionary – Dataset Introduction
- Data Dictionary – Element Description
- Data Dictionary – Unique Identifier

Files and Links:

- ✗ [DataDictionary_Introduction.csv](#)  ☒ Mark as main file 
- ✗ [DataDictionary_ElementDescription.csv](#)  ☐ Mark as main file 
- ✗ [DataDictionary_UniqueIdentifier.csv](#)  ☐ Mark as main file 
- ✗ [Bonga_Data_BWT.csv](#)  ☐ Mark as main file 
- ✗ [Bonga_Data_SMWT.csv](#)  ☐ Mark as main file 
- ✗ [Horro_Data_BWT.csv](#)  ☐ Mark as main file 
- ✗ [Horro_Data_SMWT.csv](#)  ☐ Mark as main file 
- ✗ [Menz_Data_BWT.csv](#)  ☐ Mark as main file 
- ✗ [Menz_Data_SMWT.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 on which the data collection starts
- End_Date: The date on 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

This is the basic structure, but the tab is customizable according to the needs of the author (i.e. adding specific sections about location, methodologies and additional notes).

Data Dictionary – Element Description

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 the element is found
- 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

Acceptable_Values
1 2 3 4 5
Black Red White
[0, 2000]

Fixed selection from multiple options, separated by vertical bar

Random range of values. The lowest and the highest between square brackets, separated by comma

- 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

Data Dictionary – Element Description

	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

When the dataset is composed by several files (and tabs), the various elements can be all listed in the same “Elements Description” file. In this case, it is good to add a row with the tab description for each dataset file (Bonechi 2018).

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

The General Dataset Curation Guide sets a standards, but it is still a work in progress. Certain types of files may need *ad hoc* solutions (i.e. plot layout data, or big files about genetic data, using different file format).

After today's exercise, each one of you can start applying it in his own work, until it becomes common practice. You can also signal potential issues to the data curation staff, helping to expand the future version of the guide.

Layout Marchouch 2013-14									
40	15160	→	15141						
39	15121	→	15140						
38	15120	→	15101						
37	15081	→	15100						
36	15080	→	15061						
35	15041	→	15060						
34	15040	→	15021						
33	15001	→	HIBTT-14-5	15020					
32	14160	→	14141						
31	14121	→	14140						
30	14120	→	14101						
29	14081	→	14100						
28	14080	→	14061						
27	14041	→	14060						
26	14040	→	14021						
25	14001	→	HIBTT-14-4	14020					
24	13160	→	13141						
23	13121	→	13140						
22	13120	→	13101						
21	13081	→	13100						
20	13080	→	13061						
19	13041	→	13060						
18	13040	→	13021						
17	13001	→	HIBTT-14-3	13020					
16	12160	→	12141						
15	12121	→	12140						
14	12120	→	12101						
13	12081	→	12100						
12	12080	→	12061						
11	12041	→	12060						
10	12040	→	12021						
9	12001	→	HIBTT-14-2	12020					
8	11160	→	11141						
7	11121	→	11140						
6	11120	→	11101						
5	11081	→	11100						
4	11080	→	11061						
3	11041	→	11060						
2	11040	→	11021						
1	11001	→	HIBTT-14-1	11020					
BLOCK-5									
Planted 16-12-2013 6m x 2,5m									
40	TR			64	51				
39									
38									
37	61	→	70						
36									
35									
34									
33	HIBTON								
32	F7								
31									
30									
29									
28									
27	1	→	10						
26	30								
25	41	→							
24	40	→							
23	1	→							
22	19041	→	-19050						
21	19040	→							
20	19001	→	HIBTT-14-9	19020					
19	18041	→	-18050						
18	18040	→							
17	18001	→	HIBTT-14-8	18020					
16	17160	→							
15	17121	→							
14	17120	→							
13	17081	→							
12	17080	→							
11	17061	→							
10	17060	→							
9	17001	→	HIBTT-14-7	17020					
8	16160	→							
7	16121	→							
6	16120	→							
5	16081	→							
4	16080	→							
3	16061	→							
2	16060	→							
1	16001	→	HIBTT-14-6	16020					
BLOCK-6									
Planted 17-12-2013 6m x 2,5m									



Thanks for your attention!