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)

Part I

Curation of datasets to facilitate their use and re-use

(with Microsoft Excel)

MEL monitoring evaluation and learning

ICARDA Science for resilient livelihoods in dry areas

GEOAGRO

Established in 1977, the International Center for Agricultural Research in the Dry Areas (ICARDA) is a non-profit, CGIAR Research Center that focusses on delivering innovative solutions for sustainable agricultural development in the non-tropical dry areas of the developing world.

We provide innovative, science-based solutions to improve the livelihoods and resilience of resource-poor smallholder farmers. We do this through strategic partnerships, linking research to development, and capacity development, and by taking into account gender equality and the role of youth in transforming the non-tropical dry areas.

For more information, please visit:
Main website: http://www.icarda.org/

AUTHORS
Francesco Bonechi[1]

CO-AUTHORS
Enrico Bonaiuti[1], Valerio Graziano[1] and Jane Poole[2]

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

DISCLAIMER

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

Revision History

Version	Date	Originator(s)	Reviewer(s)	Description
1.0	01/10/2018	Francesco Bonechi	Enrico Bonaiuti, Valerio Graziano	Structure, content, layout
2.0	05/12/2018	Francesco Bonechi	Enrico Bonaiuti, Valerio Graziano	Structure, content, layout
3.0	15/01/2019	Francesco Bonechi	Enrico Bonaiuti, Valerio Graziano, Jane Poole	Structure, content, branding
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

Acknowledgments

When we talk about dataset curation, the main part is always to find people interested in sharing their datasets and to participate in the curation of these for some of those aspects that only the authors can know with certainty. For this reason, I would like to acknowledge with gratitude the ICARDA staff for their availability and cooperation, providing also useful feedbacks for achieving a better result. It was through this mutual involvement that was possible to accomplish the curation of the various datasets and improve the experience to develop this guide.

I would also like to sincerely thanks Enrico Bonaiuti, who always support and supervise me along this work providing significant suggestions to grow in this field, Valerio Graziano that through his experience assist me structuring this guide, Jane Poole for her valuable comments and revision, Harrison Njamba for his important feedbacks on null values, and Zainab Azough for advising about interesting automatic tools to facilitate the file formatting.

I am also thankful to CRP Big Data in Agriculture, CRP Grain Legumes and Dryland Cereals, CRP Livestock, MEL, and GEOAGRO for the financial support received and for making possible my presence to the “Big Data Course” in Rabat and to the “Data Curation Day” in Amman. These were very precious opportunities to improve my expertise and have knowledge exchange with the other participants on this topic.

Table of Contents

Revision History II

Acknowledgments III

Acronyms V

Introduction 1

Dataset Curation Process 1

Useful references for additional reading: 1

1. Preliminary Step 2

2. Data Calculations and Summaries 2

3. Data Table 3

4. Structuring Data in the Spreadsheet 4

4.1. Formatting Features 5

4.2. Column Headers 5

4.3. Data Entry 6

4.4. Null Values 7

4.5. Dates and Time 7

4.6. Latitudes and Longitudes 8

5. Stable file formats 9

5.1. CSV file format 9

6. Data Dictionary 10

6.1. Dataset Introduction 11

6.2. Dataset Elements Descriptions 12

6.3. Unique Identifier 13

7. Conclusions and Recommendations 14

References 15

Annex A i

Annex B – Tools and Weblinks iii

Acronyms

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

Recruited as a consultant for the Big Data Platform, Francesco Bonechi was supporting ICARDA's scientists in dataset curation across different disciplines to enable them to make their research data reusable over time. Principally, he assisted them to develop structured dataset files respecting machine-readable standards, create their own data dictionary and curate the datasets maintaining the original identity and meaning of their contents. The goal was to have good examples of curated datasets for ICARDA and produce a short guide for scientists, in order to improve the quality of datasets before uploading to MEL[3] and consequently generate a persistent link (Handle[4]) on Dataverse[5] repository.

Dataset Curation Process

Data collection and management is one of the main tasks of many research activities and results depend on the systems we put in place for this. However, "the long-term value of data can be affected, for better or worse, by how well those data are curated. Unfortunately, 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. Due to this, there is the need to review and adjust these datasets appropriately.

"Data curation activities enable data discovery and retrieval, maintain data quality, add value, and provide for re-use over time" (Munoz, 2017). Thus, it will be important for anyone to have basic knowledge of this subject to be able, during research activities, to create well curated datasets.

This short guide addresses some of the dataset curation steps including cleaning files, creating a data dictionary referencing to standard vocabulary, and conversion of files from a licensed software format (e.g. Microsoft Excel), which can be opened only using the same family of products with specific supported versions, to a stable format, like CSV, compatible with many licensed and open-source products, including statistical analysis software, ensuring that the file can be read well into the future. In this way, the dataset will last well beyond the current scope keeping its validity for research purposes.

Useful references for additional reading:

- "Data Carpentry: Data Organization in Spreadsheets Ecology lesson" (Bahlai, 2017); to structure the files based on the machine-readable standards;
- "Ag Data Commons Data Submission Manual v1.3" (USDA, 2016); to develop and manage the data dictionary of the datasets.

Useful tool to support files cleaning and conversion for machine-readability:

- "Talend Data Preparation" is one of the tested software able to automatically identify the errors in the dataset and allow for cleaning and formatting actions (link available in Annex B).

[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).

1. Preliminary Step

"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 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. Data Calculations and Summaries

"Spreadsheets are good for data entry, but it's common to use spreadsheet programs for much more than this. They are used to create data tables for publications, to generate summary statistics and make figures" (Bahlai, 2017). But this all must not be included in the standard curated dataset.

In fact, the dataset should just contain the raw cleaned data. Any additions (pivot tables, graphics, formulas, etc.) must not be there since they represent the successive analysis steps and must not affect the integrity of the dataset (Fig. 2.1 and 2.2). These additions may later be presented as supplementary material with publications or contained in reports of the research.

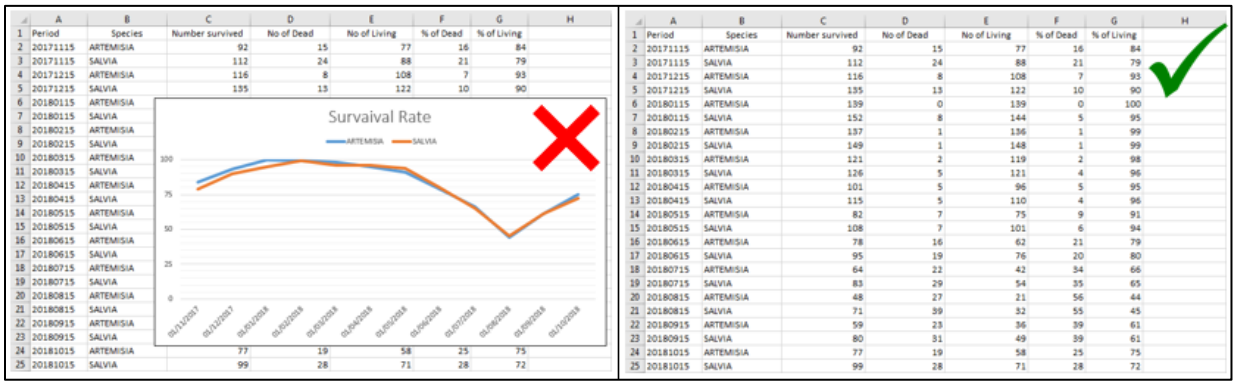


Figure 2.1. *Graphics and figures must be removed from the dataset spreadsheet tab.*

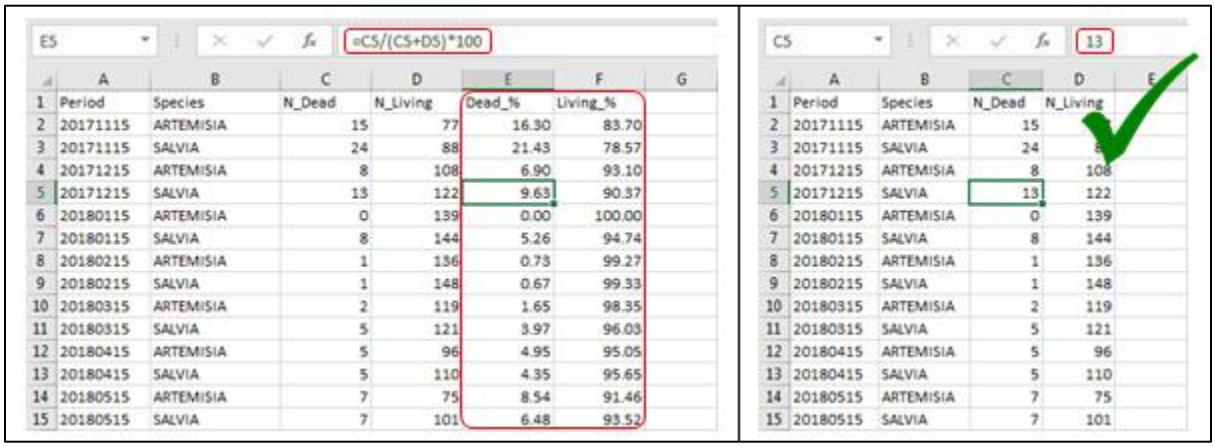


Figure 2.2. *Formulas and any other type of elaboration must be removed from the dataset spreadsheet tab. This may force to delete entire spreadsheet columns.*

Note: In some cases, the data are collected to perform specific calculations (net weight, % of production, etc.). However, even if the results of the calculations are the main part of the research activities, these should not be shown in the curated dataset. Including these may influence the use of the data by others.

3. Data Table

Another important step is the adjustment of the tables in the spreadsheet. It is not possible to have multiple data tables in one spreadsheet and use blank rows or columns to separate the data. This because the computer is not able to distinguish among the different tables creating false associations (Bahlai, 2017).

As shown in figure 3.1, the extra tables must be cut from the current sheet and pasted in a new tab. At the end of this operation, you will have more tabs in the file. Each tab will contain one of the tables previously available in the same sheet. The result is one table for each tab.

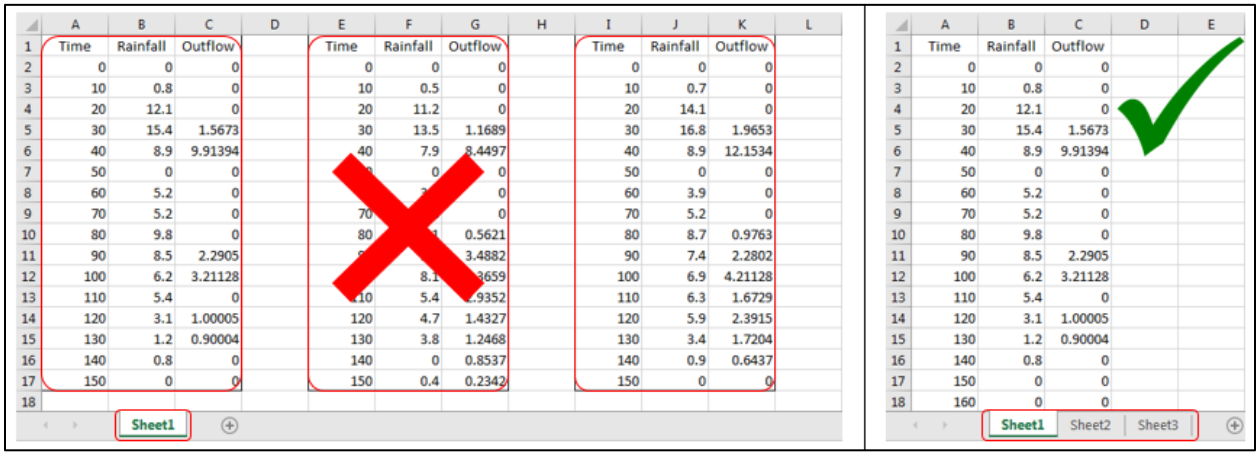


Figure 3.1. *In order to avoid false association during data system readability, blank rows and columns must not be used to separate the dataset in different tables or sections.*

Note: In this example, we create multiple tabs to organize the data file. However, “when you create extra tabs, you fail to allow the computer to see connections in the data and you are more likely to accidentally add inconsistencies in the file" (Bahlai, 2017). For this reason, if there is a link between the different sheets you should combine them into one. For example, when you create a separate tab for each day you take a measurement. This problem can be solved by adding a “Date” column, avoiding any tab repetition (Fig. 3.2).

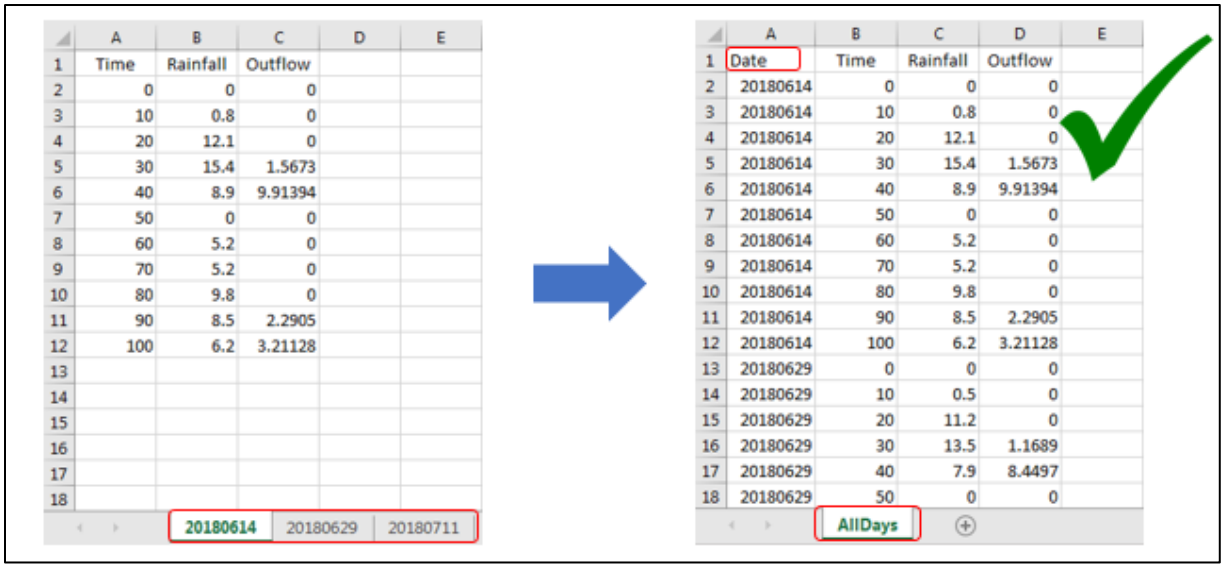


Figure 3.2. *Improving the dataset column arrangements is possible to reduce the number of tabs in your spreadsheet.*

4. Structuring Data in the Spreadsheet

Once the basic data manipulation has been carried out, we can proceed to structure the data in the spreadsheet.

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

- Put all the variables as 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).

In order to respect these principles, the data must be organized following a fixed schema, where the field (or variable) names correspond to the column header and the different observations are arranged in the related rows (Fig. 4.1).

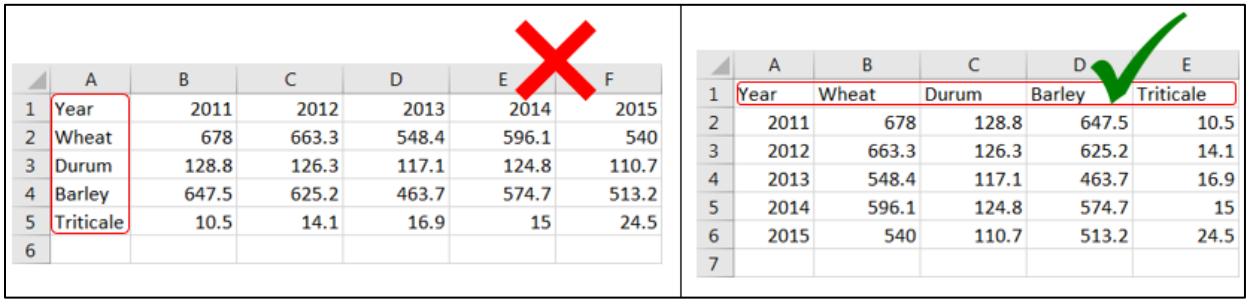


Figure 4.1. *On the left a not correct data organization. On the right the suggested data arrangement, where the columns correspond to variables, the rows to observations and the cells to values.*

Once that the dataset has been structured correctly we have to proceed to the further actions below for complete curation.

4.1. Formatting Features

The special formatting features (merged cells, borders, colors, bold, etc.) must be avoided as much as possible. All these aspects facilitate the human approach to the data but create many issues for the machine-readable processes. Each dataset should have a simple structure of columns and rows. "Consider restructuring your data in such a way that you will not need to merge cells or other aesthetic features to organize your data" (Bahlai, 2017) (Fig. 4.2).

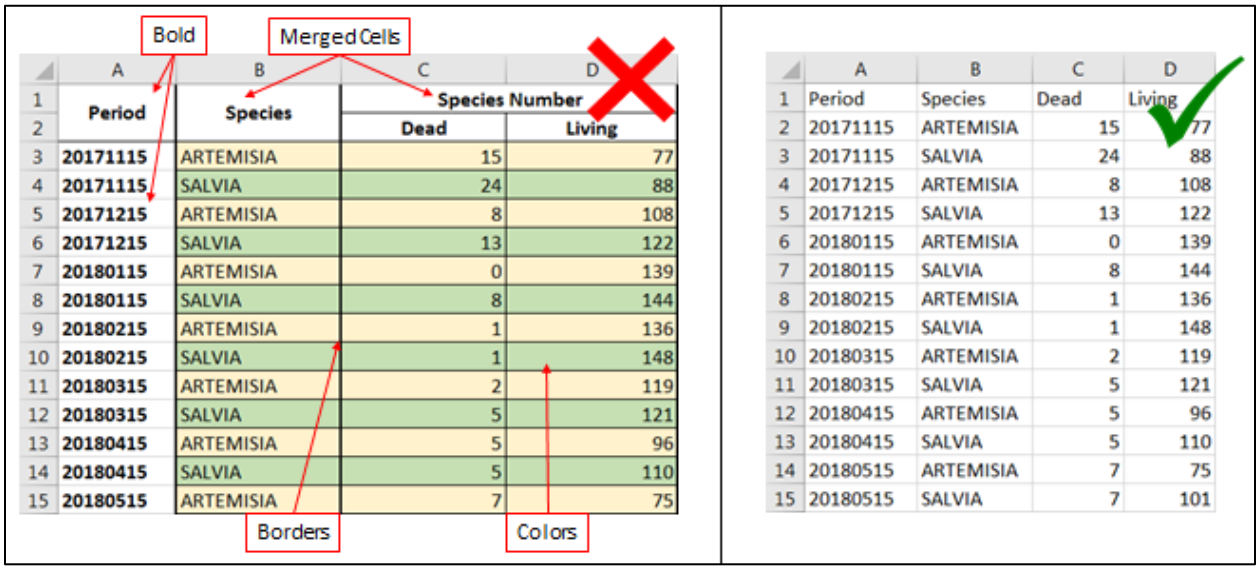


Figure 4.2. *Special formatting features are removed from the tab to facilitate the next machine-readability processes.*

4.2. Column Headers

Do not capture documentation and text descriptions in the data tables. The descriptive information can be recorded in the data dictionary or put into a "Note" column created for this purpose (USDA, 2017). Column headings should indicate the content of each column without any additional description. "Consider a limited length of your variable names. Short names can be read by most of the software and for this reason, it is suggested to use variable names that are no longer than 8 characters, beginning with a letter" (IITA, 2019). Then, ensure column headings do not contain spaces, hyphens or any other symbols. Only the underscore is allowed (Fig. 4.3).

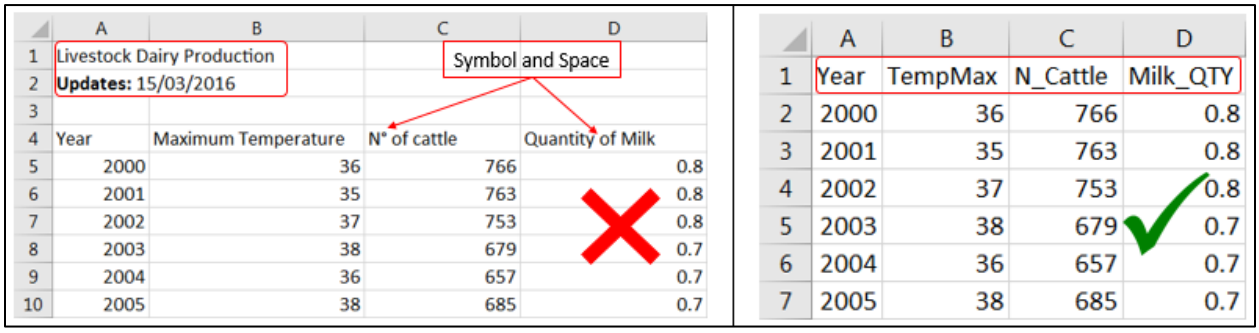


Figure 4.3. *Extra documentations are removed from the tab, and the column headers are written in a consistent way.*

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

4.3. Data Entry

Data must be entered in a consistent way always using the same code for the same value. The codes should be written carefully, taking care to write them always using the same format in terms of spaces, symbols and other characteristics adopted (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 1.4. *The same code entered without using a consistent format (spaces and symbols) on the left and the same code entered in a consistent way on the right.*

No more than one piece of information can be in a single cell. If further measurement details need to be added, they must be entered into additional columns, keeping the values separated to prevent issues in the analysis and keep clean the whole structure.

This also applies for highlighted cells and comments. Although it is common to use these functionalities to add some notes to the data, these must be removed since they may create problems for machine readability. These observations can be entered in new columns. It is also possible to create a "Notes" column to report the comments or other information (Fig. 4.5).

There is no need to include units in cells. They will be reported in the data dictionary (dataset elements descriptions). However, if many different units may be used during the data collection, and you need to enter this information in the dataset, consider adding an extra column to report this.

	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. *On the left, a set of data with comments, highlighted cells, and column "B" that contains more than one piece of information in one cell (weight and sex). On the right, the curated dataset with the additional columns to enter all the different values to facilitate dataset analysis.*

Note: When writing text in cells (as for the "Notes" column in figure 4.5), they can only contain text and spaces. This means that adding characters such as newlines, tabs and vertical tabs must be avoided (Bahlai, 2017).

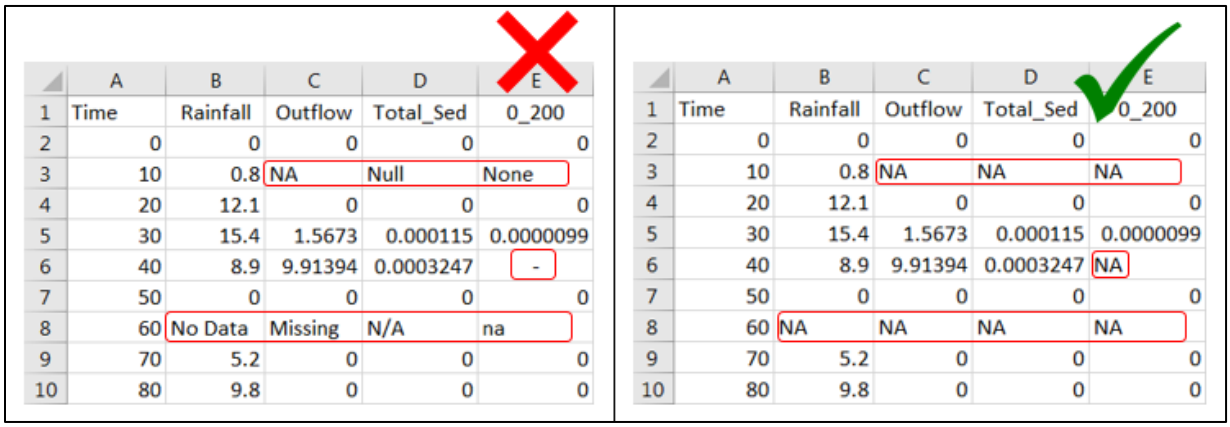
4.4. Null Values

Null values must be represented differently from "0". In fact, "0" corresponds to a measured data while null value means that the data has not been measured. By not entering the value of your observation, the computer will interpret this data as unknown or missing (null). "Because of this, it's very important to record zeros as zeros and truly missing data as nulls" (Bahlai, 2017).

Null values or missing information can be represented differently in datasets. This is because each statistical program requires a specific style to accept and read these. Therefore, how these values should be reported in the dataset will depend on the software used to analyze the data.

It is common to represent null values with blank cells, alternatively "NA" or "NULL" are good options (White, 2013). Other possibilities to indicate null or missing values is the use of numerical values (e.g. 999, -999) or other codes and text indications (e.g. Missing, No data, None, -, +), but these are not recommended since they can cause issues for the utilization of several software (White, 2013).

In any case, it is essential to select for one clear and consistent null indicator (Fig. 4.6), that must be defined in the metadata descriptions (e.g. data dictionary).



	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. *Inconsistent missing values on the left and null values correctly managed on the right.*

Note: Explicit missing value representations are better than empty fields, but this requirement depends on the software used to analyze the data. In fact, especially for date fields, dealing with null values has its own complications as most databases do not allow a null value for a date. This means that if the database that we will use to analyze the data is already known, the decision to represent missing data during dataset curation has to be made on what value is accepted from that database for null value in a date variable. However, 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).

4.5. Dates and Time

It is common that entering data as dates and hours in the spreadsheet, specific software will manage those data representing them according to some spreadsheet program default standard. This may create ambiguities in the datasets and problem for the next users. To avoid these issues, the special functionalities available must not be used since they are usually guaranteed to be compatible only within the same family of products (Fig. 4.7) (Bahlai, 2017).

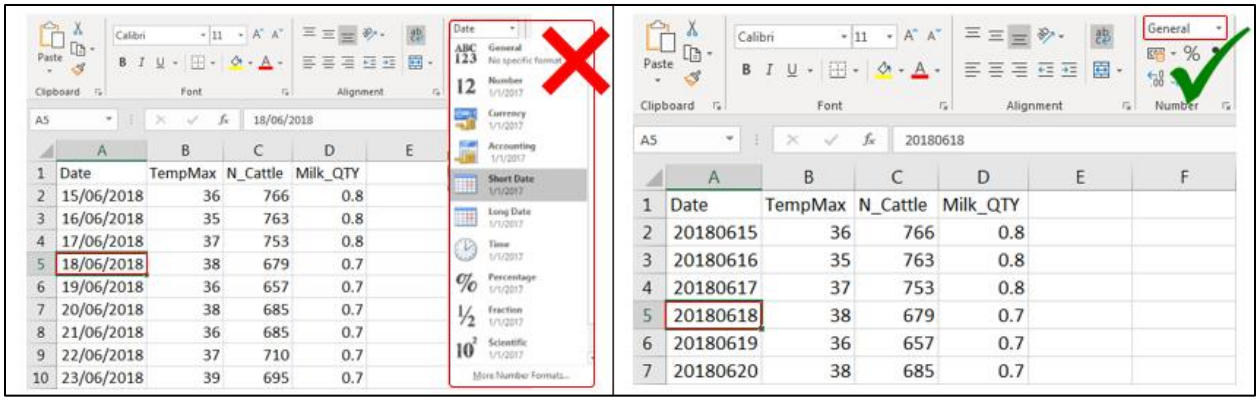


Figure 4.7. *Adoption of software specific format on the left and adoption of no software specific format on the right.*

For this reason, based 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). "For example, March 24, 2015 17:25:35 becomes 20150324172535. Such strings will be correctly sorted in ascending or descending order and by knowing the format they can then be correctly processed by the receiving software" (Bahlai, 2017).

Another option to remove any ambiguity in the dataset is to store the values of years, months, days, hours, minutes and seconds in different columns. In fact, "treating dates as multiple pieces of data rather than one makes them easier to handle" (Bahlai, 2017).

4.6. Latitudes and Longitudes

One important aspect about data collection is to track the position of where the measurements have been taken. However, in the representation of global position information, there are several ways to report latitudes and longitudes data. Among many, the recommended standard for the representation of them 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).

Thus, based on the proposed standard (Callahan, 2009):

- Latitudes is stored as numeric values in the range of [-90,90] with units of decimal degrees. Positive values indicate the Northern hemisphere while negative values the Southern one.
- Longitudes is stored as numeric values in the range of [-180,180] with units of decimal degrees. Positive values indicate the Eastern hemisphere while negative values the Western one.

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. *On the left, in bold, the suggested standard for coordinates representation. On the right, latitude and longitude expressed in decimal degrees in a set of data.*

5. Stable file formats

After completion of all the procedures mentioned before, now the dataset needs to be saved in a stable file format. When saving the files using licensed software format such as Microsoft Excel, the documents may not open using other software or even using Excel itself, if the dataset was created in an older version that is not supported anymore. For this reason, 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 files are the preferred data format for most data repositories and are recommended for publishing machine-readable tabular data. However, before proceeding with the conversion, it is important to check the Excel file tabs. This is because it is not possible to save in CSV format a spreadsheet that contains more than one tab. So, in case the Excel file of the dataset contains multiple tabs, these must be separated out into different files. For example, if the dataset Excel file contains 5 different tabs, at the end of this operation, you will have 5 different files containing one tab each. In summary, a multi-tab spreadsheet will become several files.

Now, the Excel files with just one tab are ready to be converted to CSV text-based format.

5.1. CSV file format

CSV is a text delimited file that uses a comma to separate values. It is a common data exchange format that is widely supported by consumer, business, and scientific applications (Comma-separated values, 2018). This wide applicability "means that the data in CSV format has less chance of becoming obsolete due to inaccessibility, having longer longevity than licensed file formats. In addition, CSV files are more versatile and machine-readable (computer can extract, transform and process the data)" (USDA, 2016).

In general, ease of "conversion of Excel file into CSV format, depend on the current status of the data. The final goal is to have a single spreadsheet page with a single column header row at the top of the page" (USDA, 2016). However, if the indication given in the previous sections of this guide have been followed, the conversion should be fast and easy action. In any case, here are reported some key points that need to be checked before going through the conversion process (USDA, 2016):

- Data have a single column header row to label the dataset variables.
- Avoid as much as possible commas in your document. Since the CSV delimiter is a comma, extra commas in the text can cause errors in interpreting the data.
- Data that contains multiple tables are combined into one table or separated out into different tabs (spreadsheets) and consequently various files.
- Each file is a self-contained spreadsheet with a single tab and no other extraneous information (even the blank tabs have been deleted from the file).

Once that the spreadsheet document respects these standards, proceed to save the file as a CSV.
Now, the document saved can be opened with Excel or any other software depending on the data application purposes.

Note: From now on, the dataset is represented by the folder containing all the CSV files. Due to this, enhance the folder name "entering a descriptive title including dates, locations, and specific metrics that make the dataset unique" (USDA, 2016). In the same way, ensure that all the files inside are named with a consistent and descriptive title so that it is easy to identify their data content (Hodge, 2015).

6. 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. A data dictionary provides a concise guide to understanding and using the data" (USDA, 2016). In addition, the possibility for saving this information in a separate file facilitates keeping the raw dataset clean and easy to analyze.

The data dictionary has a crucial role for data re-use processes and to correctly understand the dataset contents. In particular (Briny, 2015):

- It helps the dataset author to remember all the details about the data over time.
- 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.

Therefore, a data dictionary makes the difference between having a re-usable dataset for research purposes or not. "It is not necessarily a documentation about the data themselves but basically a documentation to give the context of understanding that data" (Briny, 2015).

In general, when the data are managed in professional databases, it is possible to automatically generate data dictionaries by the available tools in the software (e.g. look-up tables). "This will provide a document that is consistently formatted and contains what is needed for others to understand your data" (USDA, 2016).
While, when data are managed in spreadsheets, text files, or comma-separated values, the data dictionary must be created manually. To support machine-readability, it is recommended to prepare the data dictionary as a spreadsheet. In case it is preferred to prepare it as a .doc or .pdf, the table in the document should be easily extractable (USDA, 2016).
A common approach when doing it manually is to create three main files, to save in CSV format, which contains three different levels of dataset information:

1. **Dataset Introduction:** Where introductory and background explanatory information is reported;
2. **Dataset Elements Descriptions:** Where the datasets fields (variables/columns) are listed with their related information;
3. **Unique Identifier:** Where the dataset elements, terms, and concepts are identified and clarified by reference link to the online resources (multilingual thesaurus, glossaries, catalogs, etc.).

Combined these elements can be referred to as the dataset 'metadata'. When using Microsoft Excel to create the data dictionary files, in order to save them in CSV format, the principles previously explained for the dataset structure need to be followed during the creation of these documents.

6.1. Dataset Introduction

The purpose of the "Dataset Introduction" tab is to explain the contents of the dataset. Here is the general information to make clear all the dataset aspects for next uses. The fields of this tab are (Fig. 6.1):

- **Description**: 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). It should also include the design elements that are important for interpreting the data (e.g. target population, stratification, sample, size).
- **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.
- **Latitude:** Latitude site coordinate, in decimal degrees (DD), using WGS84 datum.
- **Longitude:** Longitude site coordinate, in decimal degrees (DD), using WGS84 datum.
- **Author:** Dataset first author.
- **CoAuthor:** Dataset co-author(s).

	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).

Figure 6.1. *"DataDictionary_DatasetIntroduction" fields. A "Notes" column can be added where to specify dataset relevant features, like the code adopted to express null values (or missing information).*

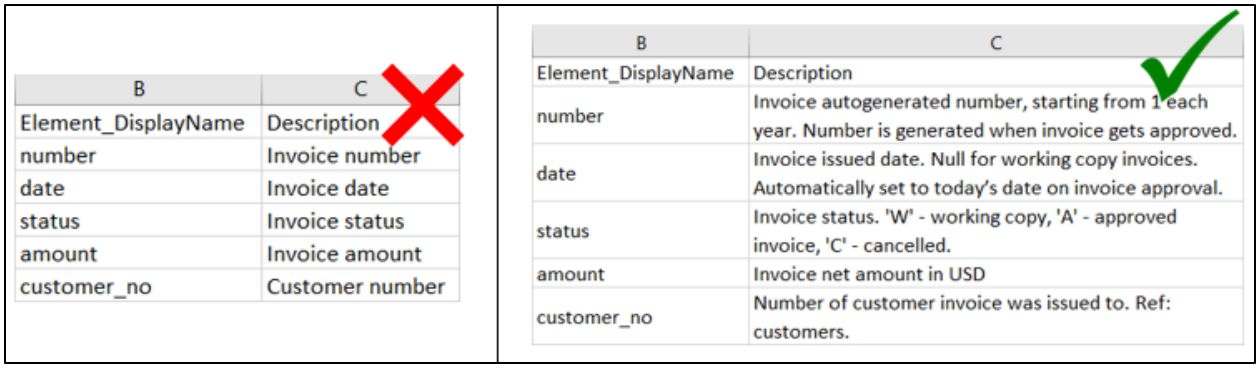
Shown above are the basic and suggested fields for the "Dataset Introduction" tab. However, if there is more information that needs to be reported (Author ORCID identifier, etc.), additional columns can be added to enrich and complete the form (CGCore, 2019). This is a versatile template that can be adapted to the various dataset needs and themes.

Please, note that the "Dataset Introduction" file allows reporting the coordinates for one location only. So, when the data have been collected in several places, the site specifics (at least latitude and longitude) should be available in the dataset itself. If not, the best option would be to create a "Site Summary" file where to report the details of the various areas (Annex A, figure A.1).

6.2. Dataset Elements Descriptions

Once the dataset introduction is complete, it is the time to explain in detail the dataset elements.
This is the core of the data dictionary since it is the document that allows the dataset users to fully understand the contents of the dataset, including the parameter names, units of measure, formats, and definitions of coded values (ORNL DAAC, 2018).
The suggested template for structuring manually the "Dataset Elements Descriptions" includes the following fields (USDA, 2016):

- **Spreadsheet_Tab:** If the dataset has multiple tabs, here is identified the name of the tab where is available the element described.
- **Element_DisplayName:** The dataset element name that is just described.
- **Description:** "A brief and complete element definition, stated in the singular, that could stand alone from other elements definitions" (USDA, 2016). It is important that descriptions are meaningful, avoiding the text holding zero 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 containing zero information on the left and descriptions with useful information on the right (Source: Kononow, 2017).*

- **Unit:** The unit of measurement adopted for the element.
- **Data_Type:** The type of data values contained in the field (e.g. date, decimal).
- **Character_Length:** The length of data values contained in the field. "For example, the maximum length for Excel is 255, so indicate 255 or less" (USDA, 2016).
- **Acceptable_Values:** The list of acceptable values in this field. The symbols adopted ("|" to separate values, "[a,b]" for the range, etc.), are based on the ISO standards (ISO 80000-2:2009, 2009).
- **Required:** Express the requirement of values in the field for dataset status and validity. It is indicated by y (yes) or n (no). In case of yes, null values are not accepted for this field in the dataset.
- **Accepts_NullValue:** Express the possibility of null values in the corresponding dataset field. This is required to run calculations on the data. It is indicated by y (yes) or n (no). In case of yes, null values are accepted for this field in the dataset.

The "Dataset Elements Descriptions" is generally based on the study of the dataset variables (column headers), by which is possible to go through their meanings and elements included. So, using this structure, the "Element_DisplayName" corresponds to the column headers names. In this way, as shown

in figure 6.3, if there are 4 columns in one tab, the dataset dictionary will have at least 4 rows corresponding to the 4 dataset tab columns.

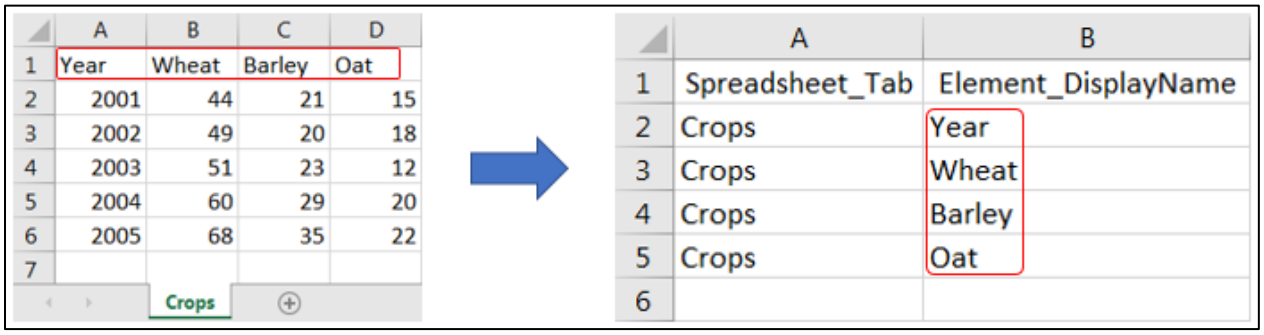


Figure 6.3. *Column headers arrangements in the "DataDictionary_ElementsDescriptions" spreadsheet.*

When the dataset is composed of several files (and tabs), the various elements can be all listed in the same "Dataset Elements Descriptions" file. In this case, it is good to add a row with the tab description for each dataset file (Fig. 6.4).

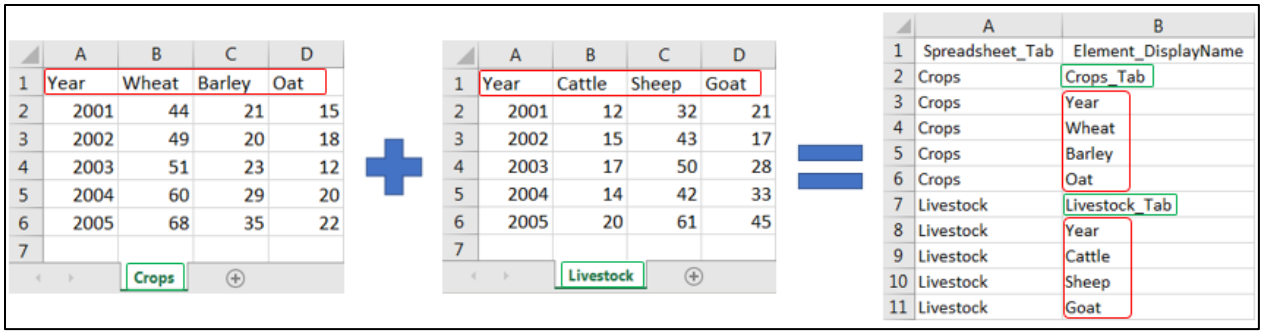


Figure 6.4. *Column headers and tab description rows arrangement in the "DataDictionary_ElementsDescriptions" spreadsheet.*

A complete example of "Dataset Elements Descriptions" tab is shown in Annex A (Fig. A.2, A.3, and A.4).

6.3. Unique Identifier

To make sure to solve any possible ambiguity, in the "Unique Identifier" tab are reported the corresponding links for the dataset terms and concepts to the online resources (multilingual thesaurus, glossaries, catalogs, etc.). This is very useful to avoid any misunderstanding on the elements analyzed and reported in the set of data (plant species, animals, etc.).

This file is structured by the following fields (Fig. 6.5):

- **Spreadsheet_Tab:** If the dataset has multiple tabs, here is identified the name of the tab where is available the element described.
- **Element_DisplayName:** The dataset element name identified. All the dataset elements (values, titles, etc.) can be identified to solve any possible ambiguity.
- **Unique_Identifier:** Where the reference link or most preferable the URIs from online resources (e.g. multilingual thesaurus) are reported. A Uniform Resource Identifier (URI) is a unique identifier that makes content addressable on the Internet by uniquely targeting items (Rouse, 2014).

- **Source:** The online resources name adopted to identify the URIs or other links of reference (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://lod.nal.usda.gov/nalt/16204	USDA	
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://lod.nal.usda.gov/nalt/41690	USDA	
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. *Representation of a standard "DataDictionary_UniqueIdentifier" file. In case of need, the "Notes" column can be added (Source: Khawam, 2017b).*

7. Conclusions and Recommendations

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 to facilitate the next uses. Following the practices described in this document, starting from a dataset in Excel format, we end up with multiple files in CSV format. Therefore, the result of this work is a folder containing the data dictionaries files and the datasets in CSV format. Other items (e.g. site relevant pictures) can be added to this in order to keep all the dataset material together.

However, it can be expensive, in terms of money and time, to deal and take care of the dataset curation processes when the data were collected, managed and analyzed a long time ago. For this reason, the greatest recommendation is to take care of all these aspects from the first data collection steps; so that these practices become an integral part of the 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/

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. (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/

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

Annex 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. *Example of "Site_Summary" file. Please, note that each time that a new file is developed, the column header descriptions and details must be reported in the "DataDictionary_ElementsDescriptions" tab.*

	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 (left) and A.3 (right). *Image of two curated datasets. "WorldBank_Education" dataset tab on the left and "INS_StudentsEducation" on the right (Source: Khawam, 2017a).*

	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

Figure A.4. *"DataDictionary_ElementsDescriptions" of the two curated datasets shown in the pictures A.2 and A.3 (Source: Khawam, 2017a).*

Annex 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