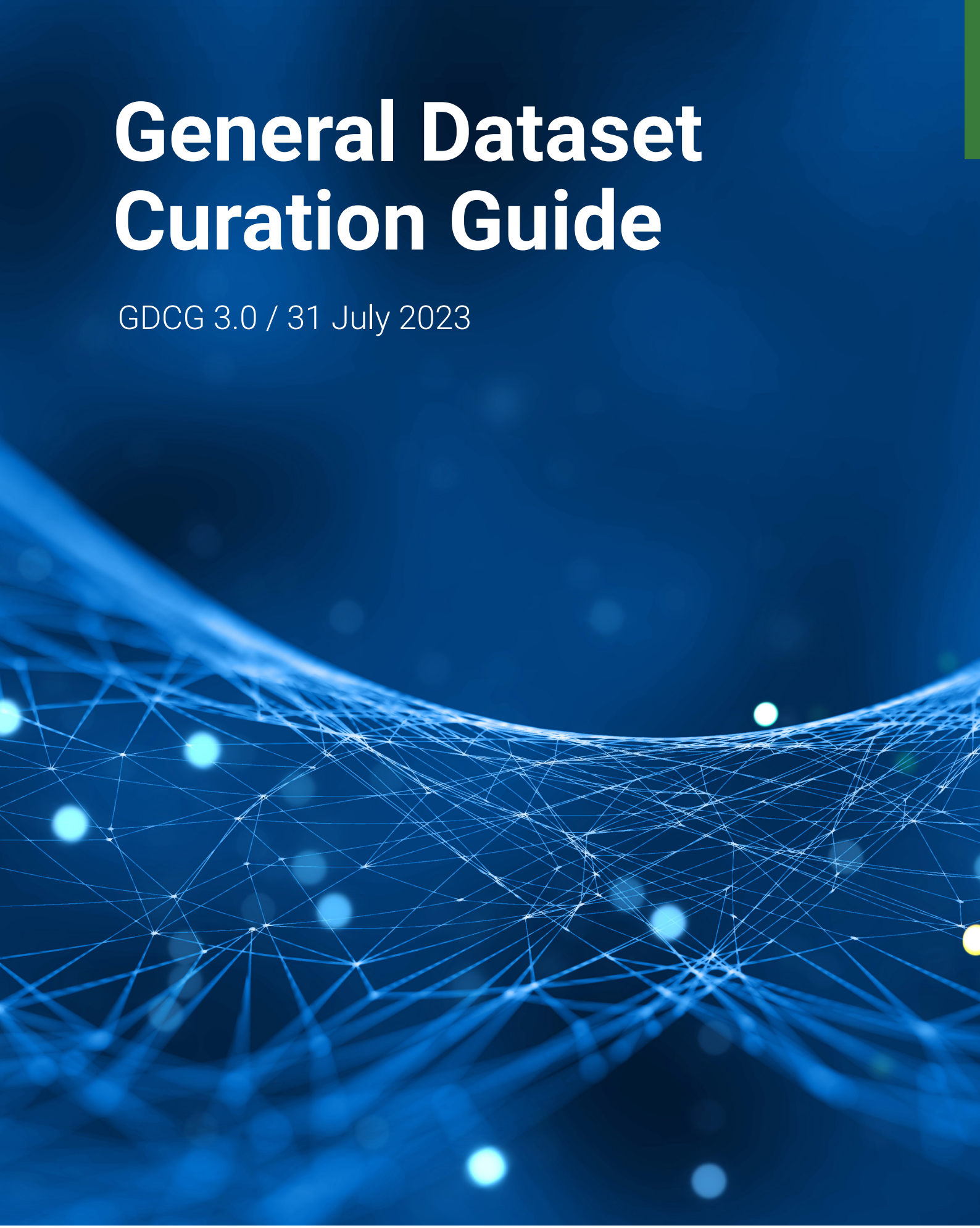
General Dataset Curation Guide

GDCG 3.0 / 31 July 2023



CGIAR

ICARDA
Science for resilient livelihoods in dry areas

CGIAR

MEL
monitoring
evaluation and
learning

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

ICARDA provides innovative, science-based solutions to improve the livelihoods and resilience of resource-poor smallholder farmers. ICARDA works 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 icarda.org

ICARDA's Monitoring, Evaluation and Learning (MEL) team improves the decision-making and impact of research organizations through four areas of expertise: **monitoring and evaluation** to plan, implement and evaluate the impact of projects and programs throughout the project lifecycle; **knowledge management** to capitalize on learning, dissemination and knowledge sharing; **data management** to collect data and ensure its quality and accuracy; and **research software development** to develop digital applications in support of the work.

For more information, please visit:
Main website mel.cgiar.org

AUTHORS
Pietro Bartolini

CO-AUTHORS
Asma Jeitani[1], Enrico Bonaiuti[1], Valentina De Col[1], Valerio Graziano[1], Sara Jani[1]

SUGGESTED CITATION
Pietro Bartolini, Asma Jeitani, Enrico Bonaiuti, Valentina De Col, Valerio Graziano, Sara Jani (2023). General Dataset Curation Guide 3.0 (GDCG 3.0). International Center for Agricultural Research in Dry Areas, Beirut, Lebanon.

[1] International Center for Agricultural Research in Dry Areas (ICARDA)

Revision History

Version	Date	Originator(s)	Reviewer(s)	Description
1.0	1/10/2018	Francesco Bonechi	Enrico Bonaiuti, Valerio Graziano	Structure, content, layout
2.0	5/12/2018	Francesco Bonechi	Enrico Bonaiuti, Valerio Graziano, Maria Garruccio, Jacqueline Muliro, Olatunbosun Obileye, Henry Juarez	Structure, content, layout
3.0	31/07/2023	Pietro Bartolini, Asma Jeitani	Enrico Bonaiuti, Valentina De Col, Valerio Graziano, Sara Jani	Structure, content, layout

Acknowledgments

The present work would not have been possible without the effort of Francesco Bonechi, who has developed the first version of the General Dataset Curation Guide (GDCG) in 2018, while engaged as consultant in the ICARDA MEL Team supporting the CGIAR Big Data Platform. The new GDCG builds on Francesco's work, improving it with suggestions and insights developed after three years of experience by the MEL Data Management sub-team.

To view a copy of the original GDCG, visit hdl.handle.net/20.500.11766/9400

Table of Contents

Acronyms

CRP	CGIAR Research Program
CSV	Comma-separated Value
DM	Data Management
DMP	Data Management Plan
FAIR	Findable, Accessible, Interoperable, and Reusable
GDCG	General Dataset Curation Guide
ICARDA	International Center for Agricultural Research in the Dry Areas
ISO	International Organization for Standardization
JPEG/JPG	Joint Photoraphic Experts Group
MEL	Monitoring, Evaluation and Learning
NA	Not Available/Not Applicable
PED	Pedigree file format
PNG	Portable Network Graphics
SAV	System for Analysis of Variables format
SNP	Single Nucleotide Polymorphism
TIFF	Tag Image File Format
URI	Uniform Resource Identifier
WGS84	World Geodetic System 1984
ZIP	ZIP Compressed File

[1] International Center for Agricultural Research in Dry Areas (ICARDA)

Introduction

ICARDA has always had a statistician and a librarian to administrate its repository, but not curators and a data manager. The data curation process started in 2018 with only a couple of consultants in charge of the curation of all datasets, in accordance with the principles of the newly developed General Dataset Curation Guide (GDCG). Nowadays, the Data Management (DM) sub-team is composed of a coordinator and a variable number of data curators (1—4), depending on the yearly needs of the MEL platform. The DM sub-team is part of the MEL team, in charge of the entire MEL[2] platform (e.g. project configuration, online repositories, metadata analysis). Each data curator is responsible for the cleaning of the data and for the preparation of a complete data dictionary, providing explanations about all dataset elements. The coordinator then reviews the final product, updates the metadata on MEL and approves the dataset to DataverseMEL[3], generating a persistent link (handle[4]).

During the last five years, the Data Management sub team has gained considerable experience in the field, while the original GCDC has reached its limits, requiring an update. At the time of its development, the GCDC described only the curation protocol for a standard dataset composed of an Excel file, however, the DM sub-team had encountered various unconventional datasets that require description. Additionally, the GCDC needs to be harmonized with the recently published CGIAR Open and FAIR Data Assets Policy (OFDA)[5].

[2] Monitoring, Evaluation and Learning (MEL) is a web-based platform for enhancing MEL in Research for Development. MEL Platform is used for managing, monitoring, and reporting on projects, from the planning phase all the way through budgeting, risk assessment, knowledge sharing, and beyond.

[3] "Dataverse is an open-source web application to share, preserve, cite, explore, and analyze research data" (The Dataverse Project, 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] CGIAR Open and FAIR Data Assets Policy

Data Curation Process

Data collection and organization is one of the main tasks during research activities. In fact, most of the project's results depend on the good management of data. 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. Unfortunately, not-curated data can quickly become unusable if nobody report all relevant information and stores it in a stable format. "Data curation activities enable data discovery and retrieval, maintain data quality, add value, and provide for re-use over time" (Munoz, 2017).

The present guide is targeted at the members of the DM sub-team and all ICARDA scientists interested in improving their data quality. It will be important for anyone to have basic knowledge of this subject to be able, during research activities, to create well curated datasets, and to ensure data are as open as possible, always FAIR (Findable, Accessible, Interoperable, and Reusable) and managed responsibly in compliance with the OFDA Policy.

The guide describes the dataset curation process in all its steps:

→ Cleaning the file.

→ Creating a data dictionary referencing to standard vocabularies.

→ Converting the file from licensed software format, like Microsoft Excel, to a stable format, like CSV (Comma-separated Value).

→ How to handle special file format and supplementary materials.

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.

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).

While doing research, it is very common to modify and restructure a dataset several times, organizing it according to researcher temporary needs, without a clear vision of the final product. A good practice, both for data analysis and data curation, is to create a copy of the original file to work on: it avoids data losses and provides a reference to fix potential errors.

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).

It is quite common to use a spreadsheet to perform data analysis using formulas, creating pivot tables, graphics, and various other elaborations. However, there is an important distinction between a standard spreadsheet and a curated dataset ready for publication.

→ A spreadsheet is an asset for internal use among a research team, used to perform specific analysis. Without proper curation, it risks being completely unusable for future research projects, making it almost impossible to share with the public.

→ A curated dataset is a reusable asset that can be released to the public as an independent product. It should contain only the raw data, and any additional elaboration must be removed **(Fig. 1 and 2)**. It helps to preserve the integrity of the dataset, and to avoid the persistence of errors and bias (even when the data are collected to perform specific calculations, it is preferable to includes only the raw data, to avoid influencing future users).

Any elaboration may be presented as supplementary material within publications or contained in research reports.

Fig. 1. Graphics and figures should be removed from the spreadsheet tab.

As a general rule, formulas should be removed from the dataset spreadsheet tab. It is advisable to remove any other elaborations to avoid bias. Depending on the type of data, this may force the curator to delete entire spreadsheet columns. Sometimes, if the information is considered relevant, the cell value can be copied and pasted without the formulas. Still, in this case, the formula used for the calculation must be reported in the data dictionary **(see section 6)**.

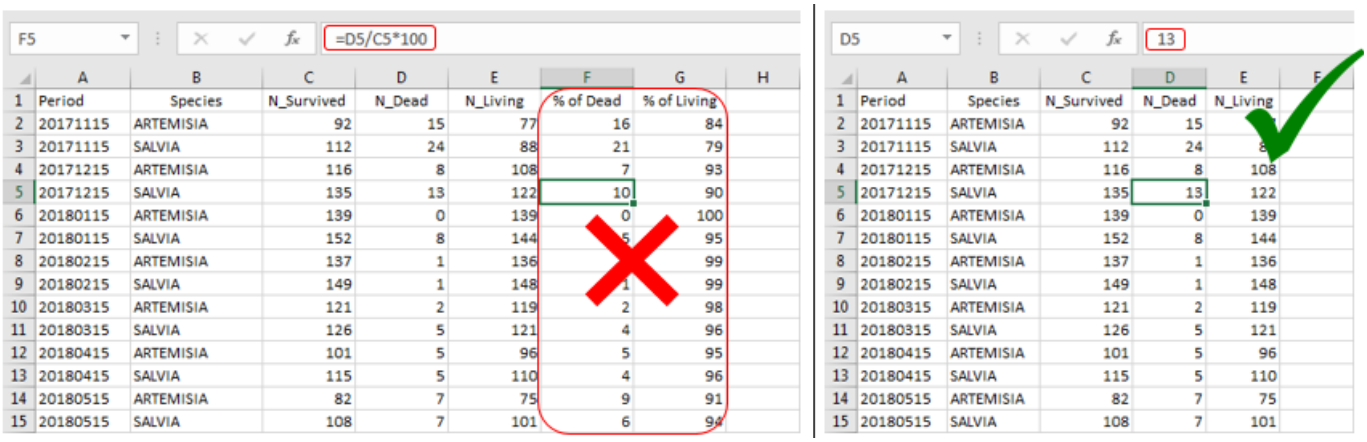


Fig. 2. Formulas should be removed from the dataset spreadsheet tab. It is advisable to remove any other elaborations to avoid bias.

3. Data Table

While performing data analysis, it is quite common to organize multiple tables in the same spreadsheet, using blank rows or columns to separate the data. This choice may help visualization and comparison but will generate additional problems because the computer is not able to distinguish among the different tables, creating false associations (Bahlai, 2017). Consequently, while curating the data for storage purposes, the table structure may require some adjustments.

If the extra tables need to be kept separated, they can be cut from the current spreadsheet and pasted in a new tab. The dataset will contain more tabs, each one containing a single table **(Fig. 3)**.

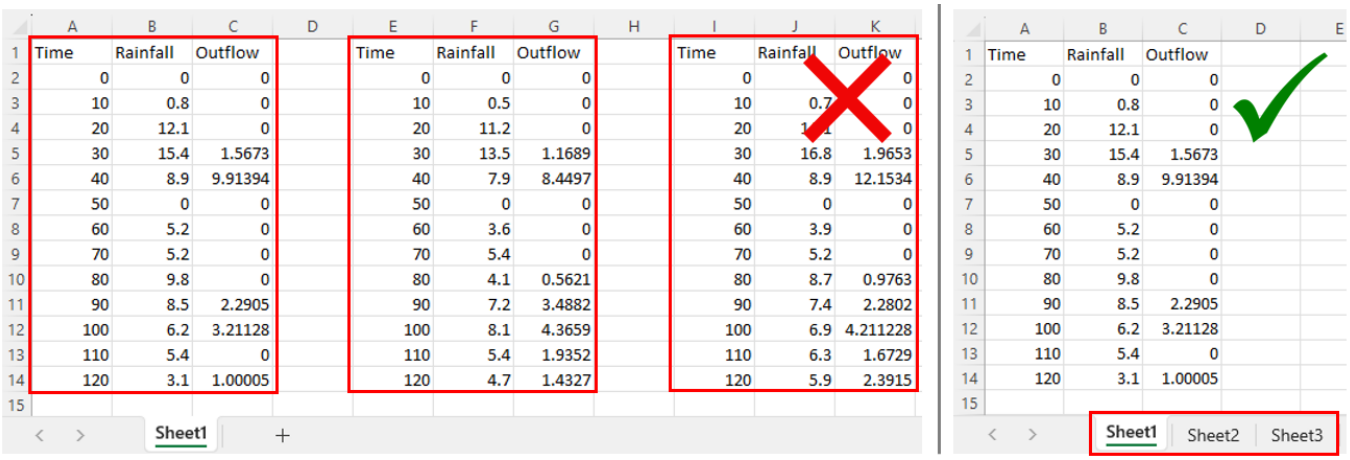


Fig. 3. In order to avoid false association during data system readability, blank rows and columns should not be used to separate the dataset in different tables or sections.

"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). If there is a link between the different tables, it would be preferable to combine them into one. To avoid confusion, the author/curator should add a distinctive variable, for example a "Date" column to identify each day while organizing a series of measurements **(Fig. 4)**.

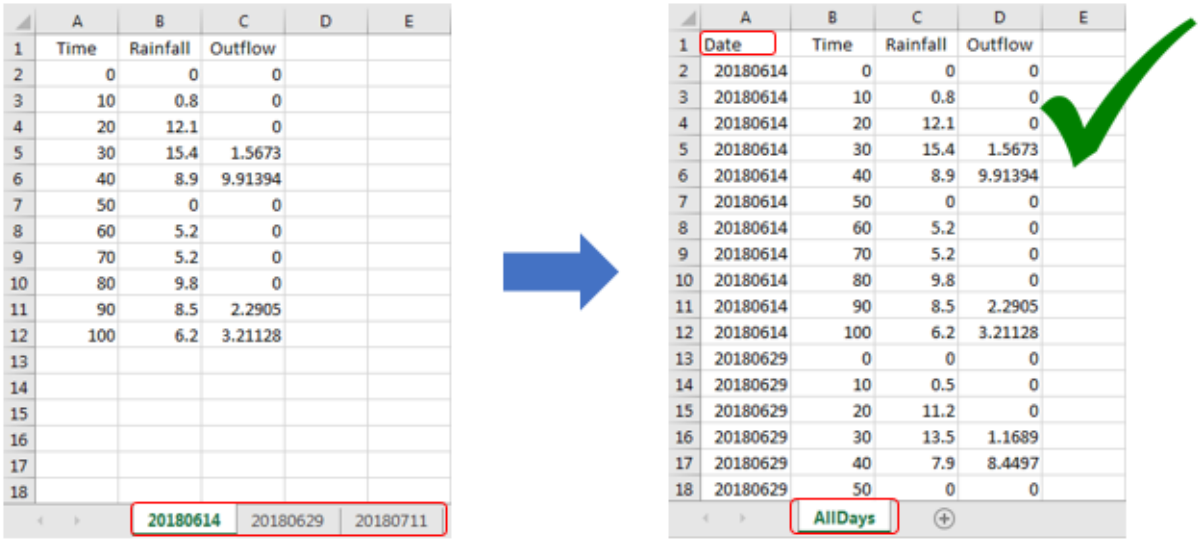


Fig. 4. Improving the dataset column arrangements can reduce the number of tabs in your spreadsheet.

4. Structuring Data in the Spreadsheet

After re-organizing the basic dataset structure, the data needs to be properly organized in each spreadsheet.

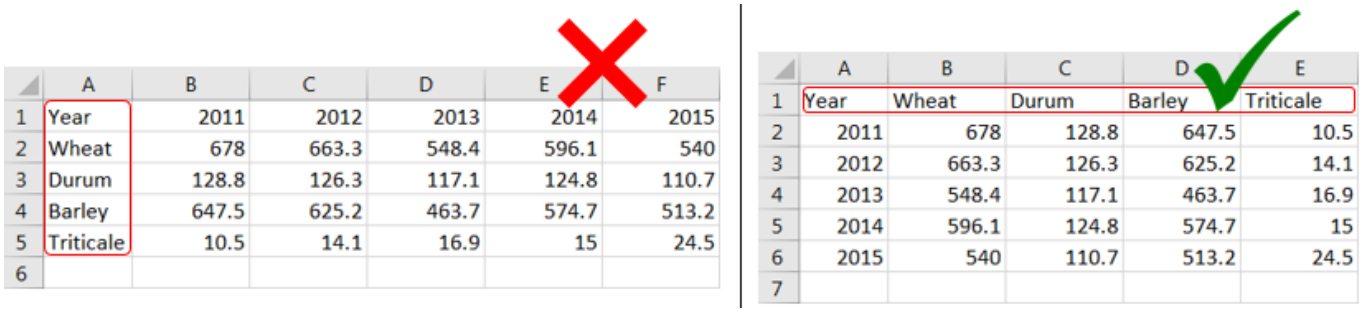
There are three main principles to follow (Bahlai, 2017):

→ 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).

To respect these principles, 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. 5)**.



	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					

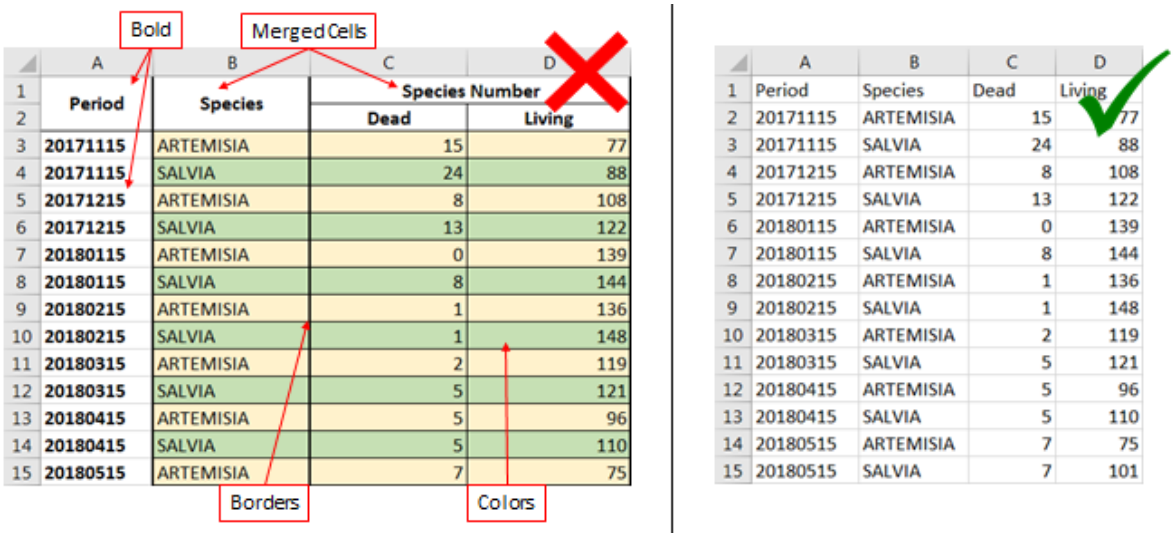
Fig. 5. On the left, a not correct data organization is shown. On the right, it is shown the suggested data arrangement, where the columns correspond to variables, the rows to observations and the cells to values.

4.1. Formatting Features

While presenting data, it is quite common to use special formatting features, such as merged cells, borders, colors, bold letters, to make them more visually pleasant, highlight certain elements, and help focus the attention of the reader. Although these features facilitate the human approach to data, they do not follow machine-readable standards, creating management and storage issues.

"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).

A curated dataset should have a simple structure of columns and rows, avoiding the use of special formatting features **(Fig. 6)**.



	A	B	C	D
1	Period	Species	Species Number	
2			Dead	Living
3	20171115	ARTEMISIA	15	77
4	20171115	SALVIA	24	88
5	20171215	ARTEMISIA	8	108
6	20171215	SALVIA	13	122
7	20180115	ARTEMISIA	0	139
8	20180115	SALVIA	8	144
9	20180215	ARTEMISIA	1	136
10	20180215	SALVIA	1	148
11	20180315	ARTEMISIA	2	119
12	20180315	SALVIA	5	121
13	20180415	ARTEMISIA	5	96
14	20180415	SALVIA	5	110
15	20180515	ARTEMISIA	7	75

	A	B	C	D
1	Period	Species	Dead	Living
2	20171115	ARTEMISIA	15	77
3	20171115	SALVIA	24	88
4	20171215	ARTEMISIA	8	108
5	20171215	SALVIA	13	122
6	20180115	ARTEMISIA	0	139
7	20180115	SALVIA	8	144
8	20180215	ARTEMISIA	1	136
9	20180215	SALVIA	1	148
10	20180315	ARTEMISIA	2	119
11	20180315	SALVIA	5	121
12	20180415	ARTEMISIA	5	96
13	20180415	SALVIA	5	110
14	20180515	ARTEMISIA	7	75
15	20180515	SALVIA	7	101

Fig. 6. Special formatting features should be removed from the tab to facilitate the next machine-readability processes.

4.2. Column Headers

Alongside the use of special formatting features, it is quite common to display descriptive information in the column headers (e.g. dates, locations, measurement units). A similar arrangement does not follow machine-readable standards and consequently any descriptive information must be removed from the data tables to be recorded in the data dictionary or moved into a "Note" column created for this purpose (USDA, 2017).

We recommend organizing column headers according to the following principles **(Fig. 7)**:

→ Column headers should be short. "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).

→ Column headers should indicate only the content of each column, without additional information. Unspecific codes (e.g. A.1, A.2, A.3) are acceptable if properly described in the data dictionary.

→ Column headers should follow machine-readable standards, without spaces, hyphens or any other symbols. Only the underscore is allowed. "Underscores (_) are a good alternative to spaces. Consider writing names in camel case (e.g. TestName) to improve readability" (Bahlai, 2017).

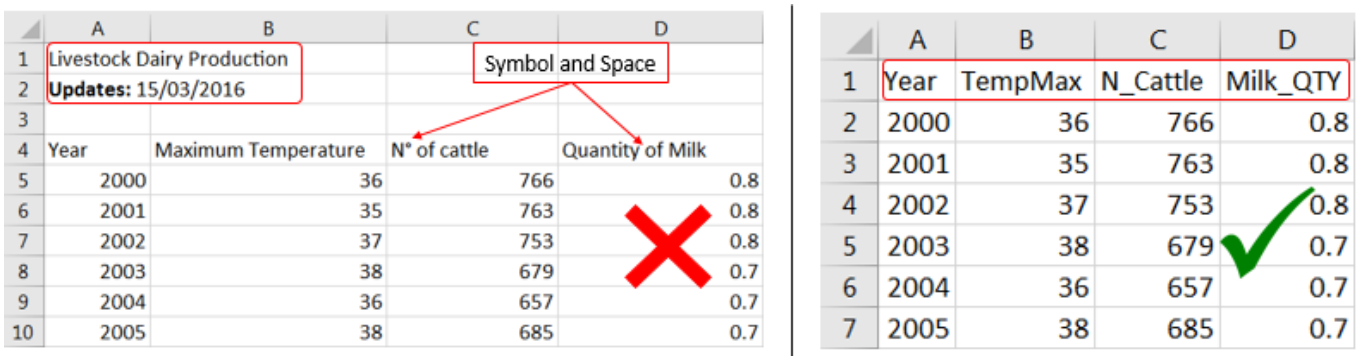


Fig. 7. Extra documentation should be removed from the tab, and the column headers should be written in a consistent way.

4.3. Data Entry

NOTE: the following recommendations are directed to data curators and authors as well. The adoption of some of the curation principles during data compilation will help curators and speed up the curation process.

We recommend managing raw data according to the following principles:

→ Data must be entered in a consistent way. The use of codes is highly recommended, but they should be written carefully, using the same format in terms of spaces, symbols and other characteristics adopted **(Fig. 8)**. The same principle stands if codes are not used, and a cell contains plain text. In this case consistent spelling (especially relevant for local names) and concise structure are recommended **(Fig. 9)**.

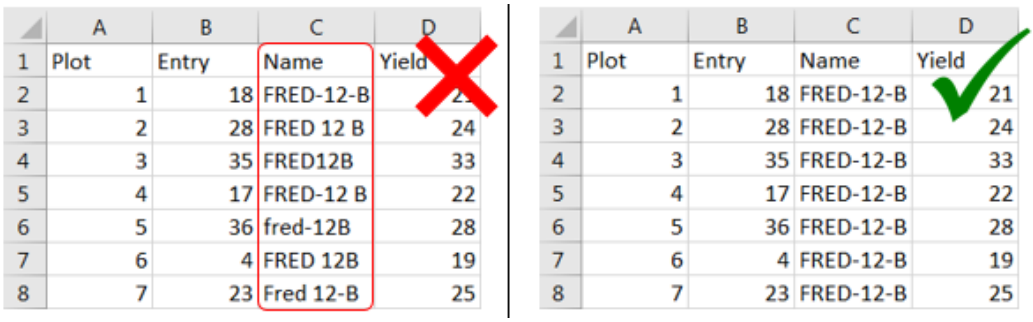


Fig. 8. The same code entered without using a consistent format (spaces and symbols) on the left and the same code entered in a consistent way on the right.

	A	B	C	D
1	Plot	Crop_2021	Change_Crop	Reason_Change
2	1	Barley	0	NA
3	2	Barley	1	Low yield
4	3	Barley	1	Poor yield
5	4	Barley	1	Low yield
6	5	Barley	1	Local varieties have lower yields
7	6	Barley	0	NA
8	7	Barley	1	Poor yield
9	8	Barley	1	Low yield
10	9	Barley	0	NA
11	10	Barley	1	Low yield
12	11	Barley	0	NA
13	12	Barley	0	NA
14	13	Barley	0	NA

	A	B	C	D
1	Plot	Crop_2021	Change_Crop	Reason_Change
2	1	Barley	0	NA
3	2	Barley	1	Low yield
4	3	Barley	1	Low yield
5	4	Barley	1	Low yield
6	5	Barley	1	Low yield
7	6	Barley	0	NA
8	7	Barley	1	Low yield
9	8	Barley	1	Low yield
10	9	Barley	0	NA
11	10	Barley	1	Low yield
12	11	Barley	0	NA
13	12	Barley	0	NA
14	13	Barley	0	NA

Fig. 9. The same "concept" expressed in different ways on the left, and the harmonized and consistent way to present it on the right.

→ No more than one piece of information should be in a single cell. If further measurement details need to be added, it is recommended to add them into additional columns, keeping the values separated to prevent issues in the analysis and keeping the whole dataset structure clean. Although highly recommended, this is not always possible. Sometimes multiple pieces of information are present in the same cell, and it isn't possible to separate them (or establish a clear priority among them). In similar cases, it is recommended to use | as separator **(Fig. 10)**.

	A	B	C	D
1	Plot	Crop	Previous_Crop	Ha
2	1	Durum wheat	Durum wheat	2
3	2	Durum wheat	Barley, legumes	1.4
4	3	Durum wheat	Barley and legumes	1.5
5	4	Durum wheat	Bread wheat	3
6	5	Durum wheat	Barley, bread wheat, legumes	2
7	6	Durum wheat	Bread wheat	4
8	7	Durum wheat	Durum wheat, legumes	1
9	8	Durum wheat	Barley, legumes	1.3

	A	B	C	D
1	Plot	Crop	Previous_Crop	Ha
2	1	Durum wheat	Durum wheat	2
3	2	Durum wheat	Barley|Legumes	1.4
4	3	Durum wheat	Barley|Legumes	1.5
5	4	Durum wheat	Bread wheat	3
6	5	Durum wheat	Barley|Bread wheat|Legumes	2
7	6	Durum wheat	Bread wheat	4
8	7	Durum wheat	Durum wheat|Legumes	1
9	8	Durum wheat	Barley|Legumes	1.3

Fig. 10. Examples about the use of separator when cell structure cannot be re-organized.

→ No highlighted cells and comment. The use of these functionalities is practical while working on the data, but it represents a problem for storage since they may create problems for machine readability. Additional observations can be entered in a newly created "Notes" column **(Fig. 11)**.

	A	B	C	D	E
1	Code	Weight_Sex			
2	20835	535M	Sister of code 20839		
3	20836	452F			
4	20837	467F			
5	20838	543M			
6	20839	458F	Sister of code 20836		
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						

Fig. 11. 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.

→ No measurement units in cells. In general, measurement units will be reported in the data dictionary (under Element description). Whenever multiple different units are used during data collection, and you need to enter this information in the dataset, consider adding an extra column to specify the measurement unit.

→ When writing text in cells (as for the "Notes" column in fig. 11), they can only contain text and spaces. This means that adding characters such as enter key and tab key must be avoided (Bahlai, 2017).

4.4. Null/Missing Values

It is common for a dataset to contain null/missing values. During data collection, it may be easy to leave empty cells to speed up the process, but these null/missing values need to be properly handled during data curation.

We recommend managing null/missing values according to the following principles:

→ Null/missing values must be represented differently from "0". While "0" corresponds to measured data, null/missing value means that the data has not been measured.

→ Explicit null/missing value are better than empty fields. It helps to maintain the dataset structure.

→ Null/missing values can be represented in different ways, according to the required style of each statistical program 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).

→ It is essential to select a clear and consistent null/missing value indicator across the dataset **(Fig. 12)**. As long as the null/missing value representation is consistent and documented, the next users can replace the choice for a null value independently, in accordance with the requirements of the software (Zwicker, 2016).

In general, we recommend adopting "NA" as a standard solution for null/missing values across MEL/Dataverse. Different solutions can be adopted for specific data types, like "genetic data" **(see chapter 7 for more details)**.

	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

Fig. 12. Inconsistent missing values on the left and null values correctly managed on the right.

4.5. Dates and time

There are different standards to register dates and time, and each data collection software manages those data according to some spreadsheet program default standards. Without proper description, this situation may create ambiguities in the dataset, making it impossible for future users to know which standard was utilized at the time of creation. Furthermore, the use of Microsoft Excel special format (e.g. "Short Date" and "Long Date") may be lost during storage. 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. 13)** (Bahlai, 2017).

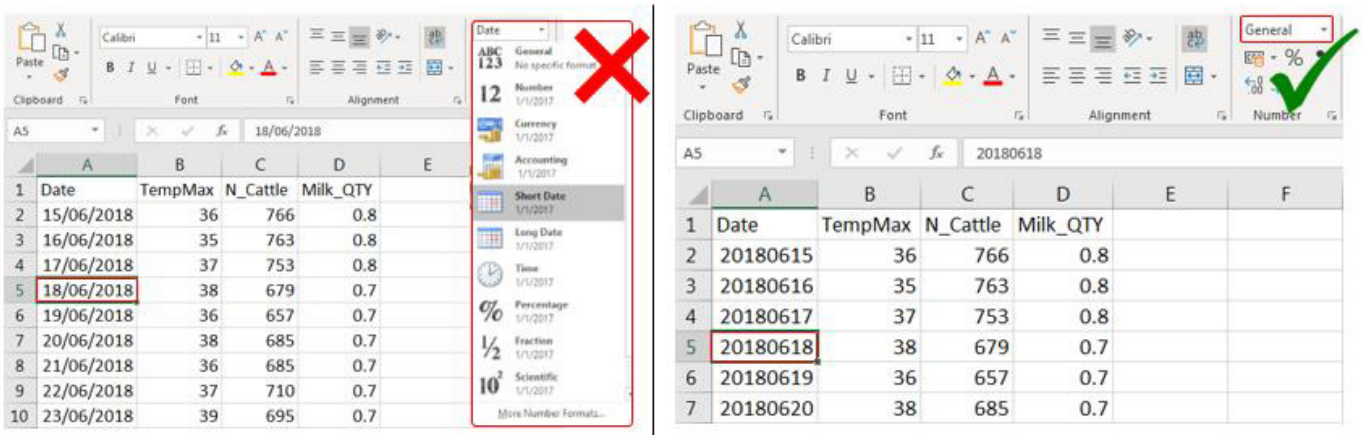


Fig. 13. Adoption of software specific format on the left and adoption of no software specific format on the right.

The recommended format to store dates and time is based on International Organization for Standardization (ISO 8601:2004): YYYYMMDD for dates, hhmmss using the 24-hours notation for time. When the two pieces of information are represented together, it becomes YYYYMMDDhhmmss. "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. Coordinates

Coordinates are an important component of data collection, allowing to track the position of where the measurements have been taken.

There are several ways to report latitude and longitude data, based on personal preferences and specific software requirements. Although they are all equally valid, 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). Consequently, any coordinates reported in a different way need to be converted during the data curation process **(Fig. 14)**.

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

→ Latitude 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 indicate the Southern one.

→ Longitude 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 indicate 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

Fig. 14. 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.

Coordinates, when associated to personal possessions (e.g. house, farmstead, etc...) are Personally Identifiable Information (PII) and may pose a security risk for the respondent. Consequently, they must be handled carefully like any other personal data (see section 4.7).

4.7. Personal data

"Data assets shall be managed responsibly, with due regard to privacy and ethical considerations in accordance with the relevant sections of the CGIAR Research Ethics Code and relevant policies on personal data protection".

According to the CGIAR Research Ethics Code, CGIAR must protect the privacy of individuals and maintain the confidentiality of PII which, alone or collected together, can lead to the identification of a particular person or household, such as:

- → Name and surname
- → Home addresses
- → Email address
- → Phone or mobile number and the advertising identifier of a phone
- → Identification (ID) card number, social security number or similar ID
- → Location data including the location data function on a mobile phone
- → Geospatial coordinates of personal or household assets, including homesteads and fields owned and/or managed or used by subjects
- → Internet Protocol (IP) address or a cookie ID
- → Any other identifier that allows for the identification of a person or a small group of persons, including people's images or voices
- → Nationality, religious beliefs or any other personal identifier, when collected together with any of the above (CGIAR System Management Office, 2020).

All these data can be grouped in three categories:

- → Direct PII: e.g. name, address, phone numbers.
- → Indirect PII: any information that allows for the identification of a person or a small group of people. The relevance of this type of information is highly context dependent, it can include, for example, jobs, religion, household assets (like houses and cars), household special characteristics.
- → Granular geo-spatial information: location data and geospatial coordinates.

In general, PII data collected by CGIAR must not be shared with third parties without explicit permission from participants. When the sharing of personal information has been authorized, the dataset should include the following consent declaration: *"Personal information including Name, Business Title, Email, Phones, Images and GPS points included in this report have been authorized in writing or verbally by the data subject"*.

Some PII may be relevant for research purposes and can be stored on secure internal repositories, however, they must not be released or made public. During data curation, all PII data must be anonymized through encoding, or removed from the dataset.

5. Stable File Format

Once curated, the dataset needs to be converted 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.

Storing the dataset using licensed software format such as Microsoft Excel, could generate problems in the future: the dataset may not open using other software or even using Microsoft Excel itself, if it was created using an older version that is not supported anymore, making it effectively unusable, and all the data practically inaccessible. For storage purposes, 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 file is the preferred data format for most data repositories, and it is recommended for publishing machine-readable tabular data. A CSV file is 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 future research purposes.

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).

"The final goal is to have a single spreadsheet page with a single column header row at the top of the page" (USDA, 2016). Following the indication from sections 2 to 4 allows a fast and easy conversion to CSV format. This is a summary of some key points that need to be checked before going through the conversion process (USDA, 2016):

→ Data must have a single column header row to label the dataset variables.

→ The use of commas must be avoided as much as possible. Since the CSV delimiter is a comma, extra commas in the text can cause errors in interpreting/converting the data.

→ It is not possible to save in CSV format a spreadsheet that contains more than one tab. Datasets that contains multiple tabs needs to be combined into one table or separated out into different spreadsheet tabs and consequently various files (for example, if the dataset Excel file contains 5 different tabs, at the end of the process the dataset will be comprised of 5 different files containing one tab each).

→ 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).

→ All the CSV files must be 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).

The creation of a separate data dictionary also helps keep the raw dataset clean and easy to analyze.

A complete data dictionary plays a crucial role in ensuring user comprehension and data reusability in the future in multiple ways:

→ It helps the dataset author(s) to remember all the details about the data over time.

→ It facilitates data sharing with collaborators, helping them to understand and use the data files.

→ It helps new users 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). Through reusability it can also improve the credit of the dataset.

The 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 data is managed in professional databases, it is possible to automatically generate data dictionary 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).

When data is 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).

The data dictionary is composed of three files in CSV format, containing three different levels of dataset information. All combined these elements can be referred to as the dataset "metadata":

→ **Data Dictionary — Introduction:** Where introductory and background explanatory information is reported.

→ **Data Dictionary — Elements Descriptions:** Where the datasets fields (variables/columns) are listed with their related information.

→ **Data Dictionary — Unique Identifier:** Where the dataset elements, terms, and concepts are identified and clarified by reference links to the online resources (e.g. multilingual thesauri, glossaries, catalogs).

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. Data Dictionary — Introduction

The "DataDictionary_Introduction" file provides general explanation about dataset content and the context of the research project. The "DataDictionary_Introduction" template used by the DM sub-team has a standardized structure, with some optional elements.

The mandatory fields of the template are the following:

→ Description: A rich and full dataset description that explains how and why the dataset was generated and informs how it should be used. Make sure that in this description are present the experiment settings (e.g. location, climatic conditions), 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 on which the data collection starts.

→ End_Date: The date on which the data collection ends.

→ Author: Dataset first author.

→ Coauthor: Dataset co-author(s).

Additional fields can be added to the template, in case more information needs to be reported (CGCore, 2019). For example, this tab may include author Open Researcher and Contributor IDentifier (ORCID), site coordinates (single location only, otherwise coordinates need to be reported inside the raw data), notes about data management.

6.2. Data Dictionary — Element Description

The "DataDictionary_ElementDescription" file is the most important component of the data dictionary, ensuring user comprehension and data re-usability in the future. It provides a detailed explanation for each dataset variable/head column, including the variable names, measurement units, formats, and definitions of coded values (ORNL DAAC, 2018).

The standard structure includes the following fields (USDA, 2016):

→ **Spreadsheet_Tab:** If the dataset has multiple tabs, the column contains the name of the spreadsheet tab where the variable/head column is present.

→ **Element_DisplayName:** The dataset element name as shown in the head column.

→ **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. 15)**.

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.

Fig. 15. Inconsistent missing values on the left and null values correctly managed on the right.

→ **Unit:** The measurement unit adopted for the element (if applicable).

→ **Data_Type:** The type of data values contained in the field (e.g. integer, decimal, text).

→ **Character_Length:** The maximum length of data values contained in the field.

→ **Acceptable_Values:** The list of acceptable values contained in this field. The symbols adopted (Pipe or vertical bar "|"to separate values, comma between brackets "[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). If 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). If yes, null values are accepted for this field in the dataset.

The structure of the "DataDictionary_ElementDescription" file is based on the raw data structure, with each row corresponding to one of the column headers. If there are four columns in the data tab, the "DataDictionary_ElementDescription" will have at least four rows corresponding to the four data tab columns **(Fig. 16)**.

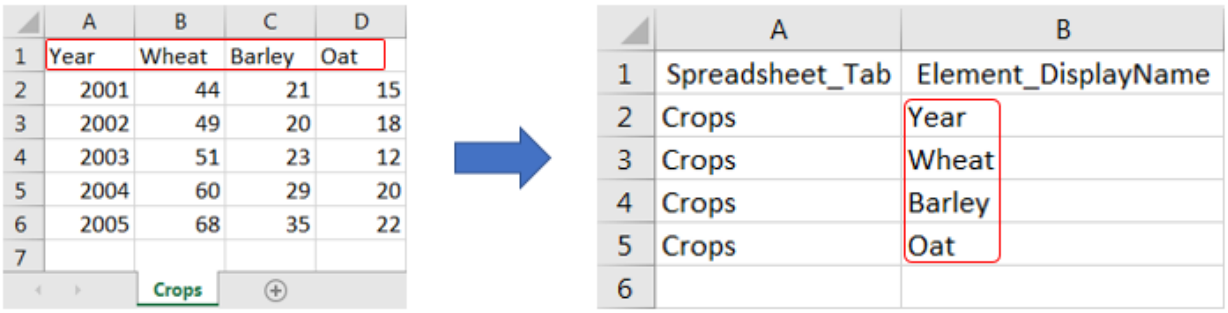


Fig. 16. Column headers arrangements in the "DataDictionary_ElementsDescriptions" spreadsheet.

Whenever a dataset is composed of several files (or multiple spreadsheet tabs), the various elements must all be listed in the same "DataDictionary_ElementDescription" file. In this case, it is good practice to add a row with a description of each spreadsheet tab **(Fig. 17)**.

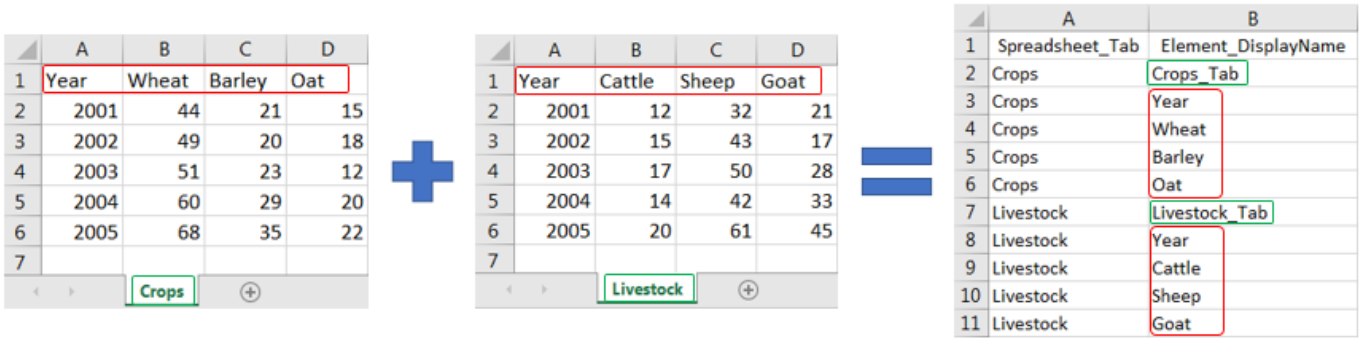


Fig. 17. Column headers and tab description rows arrangement in the "DataDictionary_ElementsDescriptions" spreadsheet.

6.3. Data Dictionary — Unique Identifier

While compiling a dataset, it is common to assume the use of certain terms to be clear, but it is not always the case, especially outside the research team. The "DataDictionary_UniqueIdentifier" file contains terms and concepts relevant to the datasets with the corresponding links to online resources (e.g. multilingual thesauri, glossaries, catalogs). This reduces any ambiguity and possible misunderstanding on the dataset's subjects (e.g. plant species, animals). We recommend listing the same terms and concepts that can be used as keywords on MEL metadata.

The "DataDictionary_UniqueIdentifier" file is composed by the following fields **(Fig. 18)**:

→ **Spreadsheet_Tab:** The column contains the name of the spreadsheet tab where the identified element is mentioned. If the identified element is mentioned in multiple spreadsheet tabs, each one of them can be listed using "|" as separator. Alternatively, if the identified element is common to all spreadsheet tabs, it can be indicated as "All_Spreadsheets".

→ **Element_DisplayName:** The name of the identified element. Potentially all the dataset elements (e.g. values, titles) can be identified to solve any possible ambiguity. While listing the same terminology is recommended, in the case of local/colloquial names we suggest listing them alongside scientific names (e.g. Hessian fly (Mayetiola destructor)).

→ **Unique_Identifier:** The column lists reference links or preferably Uniform Resource Identifiers (URIs) from online resources (e.g. multilingual thesaurus). A URI is a unique identifier that makes content addressable on the Internet by uniquely targeting items (Rouse, 2014).

→ **Source:** The name of the online resources adopted to identify the URIs or other links of reference (e.g. AGROVOC, USDA).

	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	

Fig. 18. Representation of a standard "DataDictionary_UniqueIdentifier" file. In case of need, the "Notes" column can be added (Source: Khawam, 2017b).

7. Unusual Data Format

Having curated over 300 datasets stored on MEL since 2019, the DM sub-team have encountered a few types not fitting the standard curation procedure. The following are the most common ones we were able to standardize.

Genetic data are usually stored as SNP dataset, composed by two white-space (space or tab) delimited files:

→ Pedigree file format (PED) is an original standard text format for sample pedigree information and genotype calls. The file has no header line, and one line per sample with 6 standard variables plus additional 2V, where V is the number of genotype markers.

Variable	Variable description
Family_ID	Family ID
Sample_ID	Within-family ID
Paternal_ID	Within-family ID of father
Maternal_ID	Within-family ID of mother
Sex	Sex of the animal
Phenothype	Phenotype of the animal

→ MAP is a variant information file accompanying a PED file. The file has no header file, and contains information about each marker form the PED file with the following 3-4 variables:

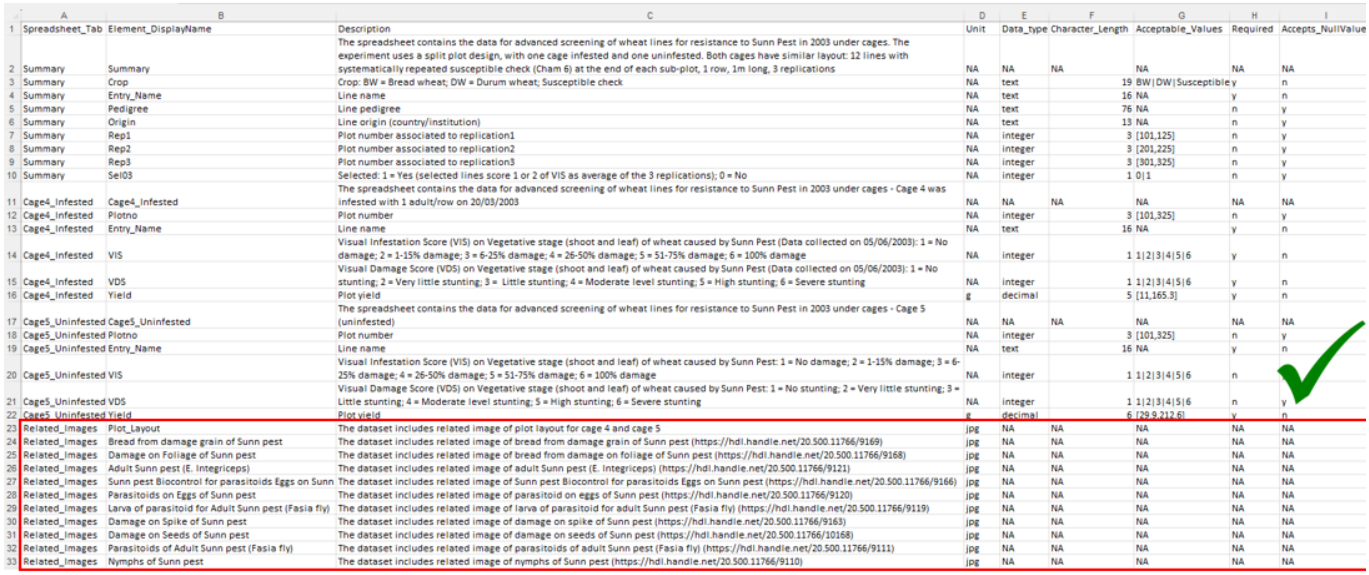
Variable	Variable description
Chromosome	Chromosome number: the value 0 corresponds to unmapped/unplaced
rs	Accession number referring to a specific SNP (single nucleotide polymorphism)
Genetic_Distance	Genetic distance between markers
Base-pair_Position	Base-pair position

This type of file is not unusual per se, and it is well known to experts of the subject. However, even if it may need no further explanation, without a data dictionary, it does not fit all the requirements of a curated dataset on MEL repository. Since this format cannot be modified without losing its functionality, the recommended curation procedure for genetic data is to prepare a separated data dictionary, describing the standard structure of a PED/MAP file, while storing the original data in a ZIP folder.

A **Photo-dataset** is composed of a series of photos/images (usually in JPEG/JPG or PNG format), with or without additional information. The recommended curation process is to organize the data in a spreadsheet, assigning one line per photo, and registering its ID. If other information is available (e.g. date of creation, location), it is possible to add how many variables as needed. Once the data is organized, a standard data dictionary can be prepared. The management of this type of dataset is generally quite simple, although it could be time-consuming. Depending on the number of photos, they can be stored as single files, one or multiple ZIP folders.

A **Data-map i**s a dataset composed of one or multiple maps (usually in TIFF, JPEG/JPG or PNG format). The recommended management is similar to the one of photo-dataset. Since each TIFF file is composed of multiple file formats, we recommend storing each one in a ZIP folder.

A dataset may include **Supplementary materials** of some kind (e.g. photos, field layout), in addition to raw data. This material is not data per se, but it could provide context information to the user. All supplementary material can be stored directly with the CSV file or linked from another asset already stored on MEL. We recommend listing all supplementary material at the end of the element description spreadsheet **(Fig. 20)**.



	A	B	C	D	E	F	G	H	I
1	Spreadsheet_Tab	Element_DisplayName	Description	Unit	Data_type	Character_Length	Acceptable_Values	Required	Accepts_NullValue
2	Summary	Summary	The spreadsheet contains the data for advanced screening of wheat lines for resistance to Sunn Pest in 2003 under cages. The experiment uses a split plot design, with one cage infested and one uninfested. Both cages have similar layout: 12 lines with systematically repeated susceptible check (Cham 6) at the end of each sub-plot, 1 row, 1m long, 3 replications	NA	NA	NA	NA	NA	NA
3	Summary	Crop	Crop: BW = Bread wheat; DW = Durum wheat; Susceptible check	NA	text	19	BW|DW|Susceptible	y	n
4	Summary	Entry_Name	Line name	NA	text	16	NA	y	n
5	Summary	Pedigree	Line pedigree	NA	text	76	NA	n	y
6	Summary	Origin	Line origin (country/institution)	NA	text	13	NA	n	y
7	Summary	Rep1	Plot number associated to replication1	NA	integer	3	[101,125]	n	y
8	Summary	Rep2	Plot number associated to replication2	NA	integer	3	[201,225]	n	y
9	Summary	Rep3	Plot number associated to replication3	NA	integer	3	[301,325]	n	y
10	Summary	Sel03	Selected: 1 = Yes (selected lines score 1 or 2 of VIS as average of the 3 replications); 0 = No	NA	integer	1	0|1	n	y
11	Cage4_Infested	Cage4_Infested	The spreadsheet contains the data for advanced screening of wheat lines for resistance to Sunn Pest in 2003 under cages - Cage 4 was infested with 1 adult/row on 20/03/2003	NA	NA	NA	NA	NA	NA
12	Cage4_Infested	Plotno	Plot number	NA	integer	3	[101,325]	n	y
13	Cage4_Infested	Entry_Name	Line name	NA	text	16	NA	y	n
14	Cage4_Infested	VIS	Visual Infestation Score (VIS) on Vegetative stage (shoot and leaf) of wheat caused by Sunn Pest (Data collected on 05/06/2003): 1 = No damage; 2 = 1-15% damage; 3 = 6-25% damage; 4 = 26-50% damage; 5 = 51-75% damage; 6 = 100% damage	NA	integer	1	1|2|3|4|5|6	y	n
15	Cage4_Infested	VDS	Visual Damage Score (VDS) on Vegetative stage (shoot and leaf) of wheat caused by Sunn Pest (Data collected on 05/06/2003): 1 = No stunting; 2 = Very little stunting; 3 = Little stunting; 4 = Moderate level stunting; 5 = High stunting; 6 = Severe stunting	NA	integer	1	1|2|3|4|5|6	y	n
16	Cage4_Infested	Yield	Plot yield	g	decimal	5	[11,165.3]	y	n
17	Cage5_Uninfested	Cage5_Uninfested	The spreadsheet contains the data for advanced screening of wheat lines for resistance to Sunn Pest in 2003 under cages - Cage 5 (uninfested)	NA	NA	NA	NA	NA	NA
18	Cage5_Uninfested	Plotno	Plot number	NA	integer	3	[101,325]	n	y
19	Cage5_Uninfested	Entry_Name	Line name	NA	text	16	NA	y	n
20	Cage5_Uninfested	VIS	Visual Infestation Score (VIS) on Vegetative stage (shoot and leaf) of wheat caused by Sunn Pest: 1 = No damage; 2 = 1-15% damage; 3 = 6-25% damage; 4 = 26-50% damage; 5 = 51-75% damage; 6 = 100% damage	NA	integer	1	1|2|3|4|5|6	n	
21	Cage5_Uninfested	VDS	Visual Damage Score (VDS) on Vegetative stage (shoot and leaf) of wheat caused by Sunn Pest: 1 = No stunting; 2 = Very little stunting; 3 = Little stunting; 4 = Moderate level stunting; 5 = High stunting; 6 = Severe stunting	NA	integer	1	1|2|3|4|5|6	n	y
22	Cage5_Uninfested	Yield	Plot yield	g	decimal	6	[29.9,212.6]	y	n
23	Related_Images	Plot_Layout	The dataset includes related image of plot layout for cage 4 and cage 5	jpg	NA	NA	NA	NA	NA
24	Related_Images	Bread from damage grain of Sunn pest	The dataset includes related image of bread from damage grain of Sunn pest (https://hdl.handle.net/20.500.11766/9169)	jpg	NA	NA	NA	NA	NA
25	Related_Images	Damage on Foliage of Sunn pest	The dataset includes related image of bread from damage on foliage of Sunn pest (https://hdl.handle.net/20.500.11766/9168)	jpg	NA	NA	NA	NA	NA
26	Related_Images	Adult Sunn pest (E. Integriceps)	The dataset includes related image of adult Sunn pest (E. Integriceps) (https://hdl.handle.net/20.500.11766/9121)	jpg	NA	NA	NA	NA	NA
27	Related_Images	Sunn pest Biocontrol for parasitoids Eggs on Sunn	The dataset includes related image of Sunn pest Biocontrol for parasitoids Eggs on Sunn pest (https://hdl.handle.net/20.500.11766/9166)	jpg	NA	NA	NA	NA	NA
28	Related_Images	Parasitoids on Eggs of Sunn pest	The dataset includes related image of parasitoid on eggs of Sunn pest (https://hdl.handle.net/20.500.11766/9120)	jpg	NA	NA	NA	NA	NA
29	Related_Images	Larva of parasitoid for Adult Sunn pest (Fasia fly)	The dataset includes related image of larva of parasitoid for adult Sunn pest (Fasia fly) (https://hdl.handle.net/20.500.11766/9119)	jpg	NA	NA	NA	NA	NA
30	Related_Images	Damage on Spike of Sunn pest	The dataset includes related image of damage on spike of Sunn pest (https://hdl.handle.net/20.500.11766/9163)	jpg	NA	NA	NA	NA	NA
31	Related_Images	Damage on Seeds of Sunn pest	The dataset includes related image of damage on seeds of Sunn pest (https://hdl.handle.net/20.500.11766/10168)	jpg	NA	NA	NA	NA	NA
32	Related_Images	Parasitoids of Adult Sunn pest (Fasia fly)	The dataset includes related image of parasitoids of adult Sunn pest (Fasia fly) (https://hdl.handle.net/20.500.11766/9111)	jpg	NA	NA	NA	NA	NA
33	Related_Images	Nymphs of Sunn pest	The dataset includes related image of nymphs of Sunn pest (https://hdl.handle.net/20.500.11766/9110)	jpg	NA	NA	NA	NA	NA

Fig. 19. Example of supplementary materials listed in the element description spreadsheets with and without links.

8. Conclusions and Recommendation

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 Microsoft Excel format, we end with multiple files in CSV format: three files forming the data dictionary and a variable number of files containing the raw data (excluding any supplementary material described in see section 7). Dataset types not described in the present guide may require adopting ad hoc solutions. The Data Management sub-team will take care of standardizing their curation processes and updating the guide in the future.

In general, it can be hard and expensive, in terms of money and time, to deal and take care of the dataset curation processes when the data were managed and analyzed a long time before. For this reason, the greatest recommendation is to take care of all these aspects since data collection; 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.

In addition, the DM sub-team can also help develop Data Management Plan (DMP) and draft data papers.

A DMP is an internal document developed in collaboration with the project manager that allows the DM sub-team to evaluate in advance what kind of support the project may need for the data curation/publication, to better allocate resources in the yearly workplan, and to publish any data assets in time to fit the deliverable deadline.

The DM sub-team encourages the drafting of data papers on interesting/innovative subject. A data paper is a scientific publication that describes the data and is used to share datasets, thereby improving their re-usability, and facilitating reproducibility and findability. It increases traffic towards associated research articles and data, leading to more citations and more collaborations.

References

Bahlai, C., & Teal, T., (Eds). (2017, April). Data Carpentry: Data Organization in Spreadsheets Ecology lesson (Version 2017.04.0). Retrieved from datacarpentry.org/spreadsheet-ecology-lesson

Big Data Platform Metadata Working Group, CGIAR. (2019). CG Core metadata reference guide. Retrieved from agriculturalsemantics.github.io/cg-core/cgcore.html

Briny, K. [University of Wisconsin Data Services]. (2015, January 23). Data Management: Data Dictionaries. Retrieved from youtube.com/watch?v=Fe3i9qyqPjo

Callahan, J. (2009). Standard Latitudes and Longitudes. Retrieved from 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 cgiarmel.atlassian.net/wiki/x/GoCC

CGIAR System Management Office. (2020). CGIAR Research Ethics Code. Montpellier: CGIAR System Management Office. Retieved from hdl.handle.net/10568/113003

Hodge, A. (2015). File Naming Best Practices.Retrieved from 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 iso.org/standard/31887.html

Khawam, H., & Najjar, D. (2017a). Statistics on Gender and Education in Tunisia. Retrieved from 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 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 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 guide.dhcuration.org/faq

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

Rouse, M., & Wigmore, I. (2014). Definition: unique identifier (UID). Retrieved from 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 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 data.nal.usda.gov/book/export/html/2769

United States Department of Agriculture (USDA) (2021). Ag Data Commons. Retrieved from doi.org/10.17616/R3G051

United States Department of Agriculture (USDA). [National Agricultural Library]. (2017, August 9). ADC 18 — Convert data files to CSV format.
Retrieved from 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 doi.org/10.7287/peerj.preprints.7v2

Wikipedia contributors. (2018). Comma-separated values. In Wikipedia, The Free Encyclopedia. Retrieved December 5, 2018, from 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 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 my.usgs.gov/confluence/pages/viewpage.action?pageId=559852026

The Dataverse Project. (2019). About the Project. Retrieved from dataverse.org/about

CGIAR

ICARDA
Science for resilient livelihoods in dry areas

CGIAR

MEL
monitoring
evaluation and
learning