Making Dataset Interoperable and Machine Readable

Operative Workshop

17 January 2019
Amman, Jordan

Led by

ICARDA
Science for resilient livelihoods in dry areas

With the Participation of

ILRI
INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

Established in 1977, the International Center for Agricultural Research in the Dry Areas (ICARDA) is one of the 15 centers supported by CGIAR. ICARDA's mission is to improve the livelihoods of the resource-poor in dry areas through research and partnerships dedicated to achieving sustainable increases in agricultural productivity and income, while ensuring efficient and more equitable use and conservation of natural resources.

ICARDA has a global mandate for the improvement of barley, lentil, and faba bean, and serves the non-tropical dry areas for the improvement of on-farm water use efficiency, rangeland, and small ruminant production. In Central Asia, West Asia, South Asia, North Africa, and sub-Saharan Africa, ICARDA contributes to the improvement of bread and durum wheats, kabuli chickpea, pasture and forage legumes, and associated farming systems. Using a systems approach, it integrates improved crop varieties with improved land and water management, diversification of production systems, and value-added crop and livestock products. ICARDA backs agricultural research with social, economic, and policy research to better target poverty and enhance the uptake of improved technologies and practices. Finally, national capacity building is the foundation stone of all ICARDA's partnerships with countries to ensure sustained agricultural development of dryland communities.

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

AUTHORS
ICARDA[1]

SUGGESTED CITATION
ICARDA (2019). Making Dataset Interoperable and Machine Readable: Operative Workshop. International Center for Agricultural Research in Dry Areas (ICARDA), Amman, Jordan.

DISCLAIMER

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

Revision History

Version	Date	Originator(s)	Reviewer(s)	Description
1.0	26th February 2019	Valerio Graziano	Enrico Bonaiuti, Francesco Bonechi	Structure, content

Table of Contents

Abbreviations and acronyms

AReS	Agricultural Research e-Seeker
BDP	Platform on BigData
CRP	CGIAR Research Program
FAIR	Findable, Accessible, Interoperable, Reusable
GDCG	General Dataset Curation Guide
GLDC	CRP on Grain Legumes and Dry Cereals
ICARDA	International Center for Agricultural Research in Dry Areas
ILRI	International Livestock Research Institute
LIVESTOCK	CRP on Livestock
OA	Open Access
OD	Open Data

Introduction

Data curation has become a key component to ensure accuracy of research and usability of its results. The International Center for Agricultural Research in Dry Areas (ICARDA) with the participation of the International Livestock Research Institute (ILRI) has led this operative workshop with the aim of a) raising awareness of the importance of datasets cleanliness for **machine reading** and **knowledge sharing** for reusability and collaboration in research, b) analyze together with the scientists their constraints in dataset compiling and use in the perspective of **promoting** their **results** by also promoting their datasets.

The **General Dataset Curation Guide (GDCG)** – compiled in partnership with the CRP on Grain Legumes and Dryland Cereals (GLDC), CRP on Livestock (LIVESTOCK), Platform on Big Data (BDP) and ICARDA GeoAgro – has been presented and the solutions it offers have been tested live by the scientists, providing ground for future discussion and more in-depth analysis of the subject.

Objectives of the workshop

The main objectives of the operative workshop were:

a) To improve the understanding of data reporting for CRPs and other donors, collecting feedback of reporting constraints;

b) To learn and share experiences of dataset curation to increase the overall quality of products before submission to curation teams and before reporting;

c) To share knowledge among scientists about data collected and plans for its dissemination;

d) Produce recommendations for management to optimize data management.

Thursday, January 17, 2019

Part 1: Introduction, datasets management ongoing practices and the GDCG

The session was opened by **Mr. Enrico Bonaiuti**, Monitoring, Evaluation & Learning Specialist at ICARDA, who welcomed the workshop participants and the participation of the International Livestock Research institute (ILRI), highlighting the value of partnership building for the adoption of shared standards in data quality for enhanced interoperability and F.A.I.R.ness[2] of data for knowledge sharing.

The General Dataset Curation Guide was introduced and the participants showed great interest in how their dataset could become more easily machine-readable and also intelligible by other scientists and public for better promotion of their research results.

Ms. Jane Poole (ILRI) went through the CRP on Livestock approach to Open Access and Open Data, informed by the CGIAR Open Access and Data Management Policy[3] (adopted in 2013) and the CGIAR Open Access and Data Management Implementation Guidelines[4] (adopted in 2014), as well as the recent developments in terms of data visualization determined by the partnership between ILRI and ICARDA: the Agricultural Research e-Seeker (AReS)[5], an open explorer for DSpace repositories implemented by ILRI, ICARDA and adopted by CGSPACE, MELSPACE and WorldFish DSpace repository.

Mr. Francesco Bonechi (ICARDA) has presented the GDCG[6] and related solutions as a way to foster the setting of standards for dataset compiling and curation for machine readability, reusability and easier usage from peers and public in general. During this, the participants came out with questions and observations in order to share their own experiences on this topic and to start addressing some of the issues and needs related to the specific characteristics of their data.

Part 2: Dataset curation constraints and opportunities presented by the scientists

The second part of the workshop was dedicated to the critical analysis of datasets potential as got-to resources for other scientists and added value for scientific publications.

Mr. Enrico Bonaiuti has moderated the session, enabling the individual flash talks of the scientists by asking the guiding questions:

1. Why my dataset was collected, what was the knowledge gap?
2. Where and When was collected?

---

[2] Findable, Accessible, Interoperable, Reusable: https://www.go-fair.org/fair-principles/
[3] CGIAR Open Access and Data Management Policy, https://cgspace.cgiar.org/handle/10947/4488
[4] CGIAR Open Access and Data Management Implementation Guidelines, https://cgspace.cgiar.org/handle/10947/4489
[5] Agricultural Research e-Seeker, https://cgspace.cgiar.org/explorer/
[6] General Dataset Curation Guide (GDCG): http://hdl.handle.net/20.500.11766/9400 and http://hdl.handle.net/20.500.11766/9440

3. How and Who collected the data?
4. Did I already develop a paper based on my data? Am I planning to publish one in 2019?
5. Did I receive requests for my data? How am I planning to promote?
6. Am I planning to collect new data in 2019?

Hereby are presented the highlights of the individual interventions and specific replies to the guiding questions based on the dataset presented.

Dr. Boubaker Dhehibi (ICARDA) stressed the importance of planning the collection and subsequent integration of datasets coming from various disciplines, mainly economic, technical, social and institutional. Standards in dataset formatting shall take into account this aspect for the best outcome in data management.

About the dataset presented:

q) Why my dataset was collected, what was the knowledge gap?
a) The dataset was collected in order to reach the following objectives:

i. Identify and characterize main livelihood types of smallholders in terms of their farms' biophysical and socioeconomic characteristics in Egypt
ii. Identify determinants, both common and livelihood type-specific, of farmers' adoptions of MRBT over ICARDA's studied area in Egypt
iii. Evaluate impacts of Mechanized Raised Bed Technology (MRBT) on whole farm productivity and profit, household livelihoods, irrigated community-landscape (multi-scale impacts)

The knowledge Gap:
Although a great deal of knowledge on the proven role of MRBT in improving water use efficiency given by irrigation, agronomic and economic studies, too few studies seek to understand (1) drivers affecting farmers' adoption of MRBT, (2) multi-aspects efficiency of MRBT (technically, economically and ecologically/environmentally), (3) impacts of MRBT on whole farms' performance and households' livelihoods. Proven knowledge on these issues will be essential for informing policies and development practices that aim disseminating the technology towards achieving food security, water resources saving, and thereby better resilience to climate change.

q) Where and When was collected?
a) During 2018 in Sharkia and Assiut Governorates (Egypt).

q) How and Who collected the data?
a) The data collected by using a semi structured questionnaire and it was collected by ARC-Egypt Team through a consultancy.

q) Did I already develop a paper based on my data? Am I planning to publish one in 2019?
a) Not yet. Planned for 2019.

q) Did I receive requests for my data? How am I planning to promote?

No. Therefore, I am planning to promote it through promoting the questionnaire used to collect such data, a methodological working paper that will apply such data.

q) Am I planning to collect new data in 2019?
a) Yes. The planned collected data will focus on the assessment of the economic and environmental impact of Red Palm Weevil on date palm farming system. But this will depend on the funding if the proposal we are developing.

Dr. Mounir Louhaichi (ICARDA) highlighted that promoting datasets along with journal articles and other information products containing research results calls for a revision of the current promotion strategies. The scientists would benefit from institutional support in learning the best ways to share their results.

About the dataset presented:

q) Why my dataset was collected, what was the knowledge gap?
a) Rangelands are recognized for their importance and value in providing society with valuable products and services. Uzbek and Tajik rangeland tenure systems do not reflect the tenure access and security needs of smallholders and large agro-pastoralists. In addition, obsolete water infrastructure and high transaction costs limit balanced rangeland utilization. This situation negatively affects the livelihoods of pastoralists and exacerbates rangeland degradation and high costs for communities and government. Therefore, the main objective of this project was to increase the returns from keeping sheep and goats on marginal lands in the Aral Sea basin and Fergana Valley through integrated livestock and rangeland management.

q) Where and When was collected?
a) The data were collected from both Tajikistan and Uzbekistan between 2015 and 2016.

q) How and Who collected the data?
a) We used a grazing gradient approach as a main tool to detect fine-scale changes of vegetation composition and its structure in the sandy desert rangelands of Uzbekistan and selected North versus South facing slopes in the mountainous rangelands of Tajikistan to monitor vegetation growth parameters such as plant cover, density and biomass production. Data were collected by ICARDA staff, NARS partners in both Tajikistan and Uzbekistan and a consultant.

q) Did I already develop a paper based on my data? Am I planning to publish one in 2019?
a) We would like to write a manuscript focusing on integrated rangeland-livestock production systems in central Asia.

q) Did I receive requests for my data? How am I planning to promote?
a) We have received no requests but when published, the data will be very useful for pastoralists as management strategies will be recommended concerning monitoring grazing pressures for sustainable recovery of the vegetation. The best promotion for a database is to valorize it in writing high-quality scientific papers.

q) Am I planning to collect new data in 2019?
a) At this moment there is no intention to collect more data. However, if funding becomes available, we may pursue this study further.

Dr. Peter Hloniphani (ICARDA) underlined the importance of an accurate planning that takes into account data collection at deliverable level. This measure would allow more accurate time and resources allocation to data collection and dataset production in line with the best formatting practices for reusability.

About the dataset presented:

q) Why my dataset was collected, what was the knowledge gap?
a) The data were collected to develop a toolkit for monitoring and assessing rangeland vegetation that could be rapidly implemented while retaining accuracy. Rangelands are recognized for their importance and value in providing society with valuable products and services, such as firewood and mitigating climate change. In such ecosystems, effective management is needed for sustainable plant growth and survival, as rainfall availability is unreliable, uncontrolled grazing is high, and soil nutrient status is poor. To achieve this goal, management and utilization options identified for a specific rangeland need to be holistically integrated with monitoring indicators in a manual-style decision support system for the long-term sustainable production of rangelands exposed to grazing pressures.

q) Where and When was collected?
a) The data were collected from both Tajikistan and Uzbekistan between 2015 and 2016.

q) How and Who collected the data?
a) We used a grazing gradient approach as a main tool to detect fine-scale changes of vegetation composition and its structure in the sandy desert rangelands of Uzbekistan and selected North versus South facing slopes in the mountainous rangelands of Tajikistan to monitor vegetation growth parameters such as plant cover, density and biomass production. Data were collected by NARS partners in both Tajikistan and Uzbekistan.

q) Did I already develop a paper based on my data? Am I planning to publish one in 2019?
a) A manuscript focusing on one aspect of the research (Tajikistan)- the effects slope, season and grazing on vegetation growth has been developed and will be integrated with data collected from animal monitoring in 2019.

q) Did I receive requests for my data? How am I planning to promote?
a) We have received no requests but when published, the data will be very useful for pastoralists as management strategies will be recommended concerning monitoring grazing pressures for sustainable recovery of the vegetation. The best promotion for a database is to valorize it in writing high-quality scientific papers and this is on-going.

q) Am I planning to collect new data in 2019?
a) At this moment there is no intention to collect more data. Perhaps after few years from now, it would be useful to assess the impact grazing and season, over time, on vegetation growth.

Dr. Mourad Rekik (ICARDA) stated that datasets should be promoted after the final research results have been published, not to disseminate unnecessary or incomplete data and for the dataset to function as a boost to the main information product. This position was discussed by the participants to be found careful and valid, although the necessity for this kind of approach is determined by the nature of the research and its data.

About the dataset presented:

q) Why my dataset was collected, what was the knowledge gap?
a) This is part of the genomic work within CRP Livestock screening for functional genes in livestock breeds. Under this heading, ICARDA is working on screening genes for fecundity, heat tolerance and resistance to diseases in sheep and goat breeds of East Africa, Nile Valley and North Africa.

q) Where and When was collected?
a) The data were collected from various local sheep breeds slaughtered in 5 slaughterhouses in highly infested locations with gastro-intestinal nematodes in North Tunisia during 2017.

q) How and Who collected the data?
a) The abomasa were collected and the gastric nematodes were recovered then identified at the National School of Veterinary Medicine of Sidi Thabet, Tunisia. Other biological samples (blood, feces…) were also collected on the animals immediately before slaughter. All the data was collected by staff of the National School of Veterinary Medicine of Tunisia after a clear protocol was agreed with ICARDA.

q) Did I already develop a paper based on my data? Am I planning to publish one in 2019?
a) A manuscript on the phenotypic variation among local sheep breeds in Tunisia to infestation with gastro-intestinal parasites was drafted and is now being reviewed by all co-authors prior to submission in 2019.

q) Did I receive requests for my data? How am I planning to promote?
a) We received no requests but when published, the data will be very useful for parasitologists and epidemiologists after the 5 years' embargo. The best promotion for a database is to valorize it in writing high-quality scientific papers and this is on-going.

q) Am I planning to collect new data in 2019?
a) In addition to gastro-intestinal nematodes and in the same line of thinking, a work started in 2018 regarding resistance of Tunisian sheep breeds to ticks. Collection of the field samples will continue until mid-2019 and afterwards, we should be able to compile the whole data (nearly 500 sheep were sampled) and to curate it in the same way as we did for the present one. Globally, we anticipate that all ICARDA-ENMV work on resistance of sheep and goats to parasites will be reported in the form of curated databases.

Dr. Mira Haddad (ICARDA) has noted how promoting evidence through datasets can be an effective way to strengthen the impact of research results in addition to allow a more accurate quantification of actual and potential improvements derived from the research activity. This approach is likely to attract more funds from donors because of a more transparent way to conduct research.

About the dataset presented:

q) Why my dataset was collected, what was the knowledge gap?
a) The datasets, as in the introduction folder that will be shared soon with you, are soil moisture dataset collected along a hillslope in Jordanian rangelands. The data relates with Time Domain Reflectometry (TDR) records taken in 20, 40, 60, and 80 cm soil depth increments over the soil profile. The data represents the spatial-temporal pattern on micro water harvesting intervention impacts on soil moisture covering one complete rainy season, concretely December 2017 to September 2018. The data needed to understand the relationship between soil, water, and plant in restored areas using soil-water conservation method.

q) Where and When was collected?
a) Collected in a weekly basis (1-2 readings per week) covering the period from Dec. 2017 – Sep. 2018.

q) How and Who collected the data?
a) The data was collected by using IMKO TRIME (PICO T3/IPH44) hand held Time Domain Reflectometry (TDR) device.
The person collected the data is trained Japanese students, Sayo Fukai a master student from Tottori University – Japan, who conducted a Master research for one year on "Effect of micro-catchment water harvesting on soil moisture condition in Jordan's Badia".

q) Did I already develop a paper based on my data? Am I planning to publish one in 2019?
a) Draft for the paper has been developed and intended to be published in 2019, the paper title "Rehabilitation of degraded rangelands in Jordan: impacts of mechanized micro rainwater harvesting on the hill-slope scale soil water dynamics"

q) Did I receive requests for my data? How am I planning to promote?
a) So far, our data request to be share with the project donors, we are open to share the unprocessed datasets, with limited conditions to share and copyrights for publishing.

q) Am I planning to collect new data in 2019?
a) The data is ongoing collected for by the local people who was trained to use the TDR device, data entry and validation done by Stefan.

Part 3: Live data curation

The concluding session was moderated by **Mr. Francesco Bonechi**, **Mr. Enrico Bonaiuti**, **Mr. Harrison Njamba** and **Ms. Jane Poole**, who have supported live the scientists in implementing the dataset curation practices object of the workshop on the spot. This session has been fundamental to test the curation

options proposed, absorb effective measures and discuss methodologies by confronting related limitations in comparison with the dataset curation solutions offered during the workshop.

The exercise started in this session was agreed to be completed by the scientists within the next week, implementing the dataset curation practices for enhanced effectiveness in their research activities. The scientists also agreed to follow up with their institutions for the final version of the General Dataset Curation Guide and to provide their feedback.

In the end, **Mr. Enrico Bonaiuti** thanked the participants of the workshop for the fruitful exchange of views and regarded the discussions as the beginning of a mid-term process to operationalize data curation best practices. The final thoughts from all participants about the workshop and related expectations have been collected.

Results and next steps, recommendations for management

Based on the planned objectives of the workshop, the following summary is and outline of the results and related next steps:

a) The understanding of data reporting for CRPs and other donors was clear and consolidated.

b) The scientists improved their understanding of dataset cleanliness conditions for optimally promoting their research results.

   Next step: a proposal to enhance the support to scientists on-field for better data collection and dataset curation has been advanced. Drs. Dhehibi, Rekik and Louhaichi suggested to fund cross-project a position in Tunisia to be trained in data curation in order to improve data coming from partners since 90% of data reported are collected by national partners.

c) The participants showed great interest in the current version of the General Dataset Curation Guide (GDCG).

   Next step: the participants will follow up with their institutions to ensure an optimal finalization of the Guide for future use and dissemination. A tailored version of the guide by discipline and dataset type (other than xls or csv) can be developed if funds allows.

d) In terms of recommendations for management, the scientists have advanced specific requests for strengthening the **institutional support** toward a) **national partners** for data management and b) **scientists** for communication and promotion of their results. In particular, the figure of a Data Curation Manager has emerged as close to necessary in order to plan, report and carry out the data management for research duties, also in the perspective of an optimal promotion of datasets which is a topic to be considered for the ICARDA overall Communication Strategy.

   The **Dataset Curation Manager** has been highlighted to be an optimal solution for on-field projects involving sensible quantity of data collection and requesting and intense management overall, such as the ongoing projects on small ruminants and livestock in Tunisia, ICARDA efforts mapped to CRP on Livestock. This short-cycle (1-4 months), fast-deployment figure has been suggested to work closely with the research team, managing the **data collection** and **curation** in addition to carrying out data related **M&E duties** in collaboration with MEL team and contributing to the **data promotion**. The Dataset Curation manager shall save time to the research team, provide better data quality and ensure the accuracy and timeliness of planning and reporting operations. The experience accumulated in just a year by one or more Data Curation Managers is likely to be enough to inform data curation guidelines, to be consolidated and disseminated within ICARDA and confronted with the international standards in research, in order to obtain on-point best practices informing the institutional knowledge management strategy.

Annex 1 Workshop agenda

Thursday, 17 January 2019

Time	Topic	Speaker
8.15-8.20	Introduction	Enrico Bonaiuti
8.20-8.50	CRP Livestock approach: Open Data (w/ Q&A)	Jane Poole
8.50-9.15	Data Curation Guide (w/ Q&A)	Francesco Bonechi
Data Presentation, Moderated Flash Talks		
9.15-9.30	AFESD_Database_Final_Jan2019	Boubaker Dhehibi
9.30-9.45	Socio economic survey conducted in Jordan, Tunisia and Yemen	Mounir Louhaichi
9.45-10.00	Phenotypic resistance of sheep breeds in Tunisia to infestation by gastro-intestinal nematodes	Mourad Rekik
10.15-10.30	Vegetation survey in Tajikistan	Peter Moyo
10.30.10.45	Soil moisture datasets	Mira Haddad
Coffee Break		
11.00-13.15	Data Curation in Team	Francesco Bonechi, Jane Poole, Harrison Njamba, Enrico Bonaiuti
13.15-13.30	Closing remarks	Enrico Bonaiuti

Annex 2 List of participants

##	Name	Country/Affiliation	Contact
1	Dr. Jane Poole	Research Leader – Research Method Group (ILRI)	j.poole@cgiar.org
2	Dr. Harrison Njamba	Data Systems Manager - Research Method Group (ILRI)	h.njamba@cgiar.org
3	Dr. Boubaker Dhehibi	Senior Natural Resources Economist (ICARDA)	B.Dhehibi@cgiar.org
4	Dr. Mourad Rekik	Small Ruminant Production Scientist (ICARDA)	m.rekik@cgiar.org
5	Dr. Mounir Louhaichi	Range Ecology and Management Research Scientist (ICARDA)	M.Louhaichi@cgiar.org
6	Dr. Hloniphani Peter Moyo	PDF - Rangeland Ecology and Management (ICARDA)	h.moyo@cgiar.org
7	Dr. Mira Haddad	Senior Research Assistant - Spatial Analyses and Database Management (ICARDA)	m.haddad@cgiar.org
8	Dr. Stefan Strohmeier	Associate Scientist - Soil and Water Conservation (ICARDA)	s.strohmeier@cgiar.org
9	Mr. Enrico Bonaiuti	Monitoring, Evaluation and Learning Specialist (ICARDA)	e.bonaiuti@cgiar.org
10	Mr. Francesco Bonechi	Consultant (ICARDA)	francesco.bonechi@cgmel.org
11	Mr. Valerio Graziano	Consultant (ICARDA)	valerio.graziano@cgmel.org