Quality Research Data Acquisition Guidelines

1. Background

1.1 There is always a scope for improving the quality of research data in the National Agricultural Research and Education System. These guidelines are prepared to further improve the quality of research data.

1.2 It is re-emphasised that the data collected should be verifiable and reusable and results should be replicable in scientific research.

1.3 At present, records of field experiments are usually noted in field/laboratory note books and ad-hoc methods are often used for their storage. Digitally created data (such as sequencing, data obtained from sensors, etc.) are stored as electronic files. Individual researchers generally keep the primary data with themselves. The data resides mostly in silos and not easily accessible for reuse. Reports/tables created from the analysis are available but the raw/primary data are not clearly identified for future references. Rigorous efforts need to be undertaken to devise an institutional mechanism to aggregate the data project wise by the project team.

1.4 Documentation of the recorded data needs further improvements so as to make it reusable in future. The documentation must include all details of metadata such as location, units, transformation used, expressions of derived parameters, etc. so that the data can easily be understood by the researchers for future use. Several critical parameters such as ambient temperature/lab conditions, etc. need to be recorded, wherever applicable. Experimental plot size, sampling procedures, experimental blocking, etc. also need to be reported and need to be made more contemporary. Calibration of machines/equipments/scales need to be done periodically and proper records must be maintained.

1.5 Methods that can be help in improving the status of quality research data acquisition are Good research practices, standard operating procedures, peer review and data audit.

1.6 Coordinated trials/network projects also collect similar data. However, documentation of protocols is not up-to-date and not generally made available at a common place such as websites/repositories.

Some guidelines for quality research data acquisition are given in the sequel. Researchers across NARS may be aware of these guidelines and using them in practice, but these are not appropriately documented. Therefore, the present effort is to document these guidelines. Good research practices, standard operating procedures are subject specific and need to be developed subject wise. If already available, a copy may be kept at central place.

2. Guidelines for Quality Research Data Acquisition

2.1 Maintaining the health of our research farms is important. Every year relevant soil, water, pollutants and other quality parameters should be assessed before and after every field experiment. If the experiment is long-term in nature, then data may be collected on these aspects after every 3-4 years. Environmental indicators of land, water, pollutants, etc. should be periodically taken up and assessed every year. This would be helpful for preserving natural resources. Similarly data on weather parameters during the experiment period may also be collated.

2.2 Geo-tagging of units/experimental plots should be made compulsory for all field experiments. Unique ID/RFID tags may be used for animals wherever it is possible.

2.3 Procedures for developing field maps as well as labels with barcodes for assistance in planning need to be developed.

2.4 Develop and maintain standard experimental protocols for experiment/field preparation, data collection, design to be used, sampling scheme, minimum attributes on which data to be collected, direction on how to record data on each character under investigation, units for data reporting and decimal places to be retained, acronyms to be used, documentation, broad guidelines for statistical analysis techniques to be used, etc. These protocols must include the direction on how to record data on each character under investigation and broad guidelines for statistical analysis. and these protocols should be made available along with the technical programme every year especially in All-India coordinated Research Projects/ Network Projects. These protocols must include the direction on how to record data on each character under investigation and broad guidelines for statistical analysis. These protocols may be revisited periodically.

2.5 Detailed format for recording background information about data being recorded e.g. field history, crop, variety, sowing date, latitude/longitude, soil (if possible soil family), irrigated /rainfed, harvest date, extreme events, if any, health, age, lactation stage, initial body weight of animal, etc. may be developed at Institute level depending upon the commodity being dealt with. It should include project name, Principal Investigator (PI) /Co-PIs, years/season, objectives, etc. as given in Section 2.5 of ICAR Research Data Management Guidelines.

2.6 Laboratory equipment and other tools should be periodically calibrated and records along with log books should be available at all time.

2.7 Records of animal health may be maintained for all experimental animals.

2.8 Photographs may also be taken along with observations for monitoring and other purposes.

2.9 The image/photograph data taken may be stored in raw for facilitating re-analysis. While storing images (microscope), gels, etc. the use of minimum storage space may be considered without losing resolution, wherever possible.

2.10 For experiments which are periodic and several data points are collected, hand-held devices /electronic note books may be used for recording data in Laboratory/field / during surveys and protocols for uploading them to a Central repository at SMD / Institute level may be taken up immediately. Hand held devices must have capability of working in diverse environmental conditions, able to produce spatial coordinates and should work both in offline and online mode. To minimize costs, provisions may be made so that same device may be used by more than one scientist and individuals can access only their data using fingerprint sensors or username/password, etc.

2.11 Record keeping is important for maintaining quality of the data. A unique number may be assigned to each experiment/survey and the details can be saved at a centralized/local repository. A Data Management Plan (DMP) at the project proposal stage may be included as part of peer-review. Citable accession numbers for each data set/experiment may be given.

2.12 Uniform user friendly data recording formats may be devised in consultation with representative of institutes of different SMDs of ICAR. Data recording should be in a standardized formats (such as number of decimal places, units etc.). SI units should mandatorily be used for storage and recording. It should include the minimum Information on Experiment/Data guidelines (necessary for third party use or validation).

2.13 Quick data verification system to identify and remove errors during uploading/ posting stage should be in place.

2.14 Data type, format, naming and tagging may be synchronized with standardized vocabulary, archiving and retrieving, data sharing (exclusive use vs third party access) may be included for security and safety.

2.15 A metadata framework based on ontology may be developed for long term storage and data reuse.

2.16 Data curation and verification must be done at Principal Investigator Level. Primary/raw derived/ simulated/ compiled data and meta-data (protocols, algorithms, etc.) should be sufficient to validate research findings. Each PI and Scientists (Co-PI and associates) can have login ID, with different levels of authentication. Only PI can permit addition/deletion of saved data wherever required. Unique ID for Data submitted may be created.

2.17 Central data repositories/ Research Data Management System may be developed subject/research theme-wise.

3. General Points

3.1 Sensitisation and confidence building of scientists about data protection and data access rights need to be done. Sensitisation regarding need and benefits of third party research data acquisition and storage may also be taken up.

3.2 Quality research data acquisition procedure may be included as a module in NAARM (FOCARS) so that uniform procedures are adopted when the scientists join various institutes. Short duration refresher courses may be introduced for middle level scientists/ researchers associated from SAUs. For awareness from initial stage, new course(s) on Digital Data Management may be designed and made compulsory during PG Programmes.

3.3 Every scientist should be trained in maintaining data quality and maintenance of proper records. They should be also sensitized about the importance of data quality, data integrity and long term storage. The rights, retention periods, sensitivity, accessibility of data should be made clear to all scientists. This may be introduced as part of their one month attachment in the Institute and 3 months subject matter training. This activity may be taken up through PME cell of individual institutions. PME Cell may be strengthened and its staff may be trained on priority on data acquisition and storage. Orientation of staff of PME cell on data verification, storage, data sensitivity limit, accessibility, data retention period needs attention. Data accessibility and retention may be determined by Head of Division (HoDs)/ Institute Technology Management Committee/Unit. PME can monitor whether data is submitted for each project/product in stipulated time.

3.4 There is an imperative need to strengthen the PME cell in the institutes, with additional scientific (multidisciplinary) and senior technical officers to ensure data is collected and arranged in a consistent manner. This has been a long standing demand and needs to be taken care of.

3.5 Exposure of Nodal Officer/Staff of PME Cell to institutions where digital data/ information system is operating needs attention.

3.6 At the end of every project, along with project completion report, status report of data should be included indicating information about submission of the data files to the repository. At the time of transfers/superannuation, status report of data handed over/taken over to the next person/division, etc. should be recoded at PME cell. It should be a part of workflow process. The data should go to ICAR Data Inventory Repository.

3.7 Raw data collected through surveys (exploration, surveillance, socio-economic, cost of cultivation, etc.) should be saved. Care must be taken up to protect the privacy of respondents.

3.8 For data security, high value data may be identified and should be kept in secure place. Protocols may be kept in place for data recording, modification and retrieval of such data.

3.9 Wherever feasible research data management systems may be linked with public databases (e.g. NCBI, CYVERSE, etc.) for reducing redundancy and storage cost for minimizing the storage requirements.

3.10 Additions (shown in italics) in Section 2.4 of earlier Item 2.4 of ICAR Research Data management Guidelines

Point 2.4.1: The PME Cell shall oversee that the data is collected in a consistent, systematic manner ensuring its reliability and establish a system that permits for periodic evaluation, monitoring and flexibility for recording changes. *Period for data retention in original format would vary for each type of data. It could be 10-12 years for field/laboratory experimental data of annual crops and for social survey data, a longer period for germplasm.*

Point 2.4.5:Notebooks/registers generally offer a convenient way for data recording and tracking daily progress by all team members. These should be supplemented as needed by specialized methods of record keeping suitable for specific types of data (e.g. computer files/images, photographs, gels/gelscans, chromatograms, etc.). While keeping written records, the accepted minimum standards should include:

- Notebooks/registers should be bound and accessioned (officially numbered and catalogued by the PME Cell/Institute and issued by name to the concerned person);
- Separate notebooks/registers should be maintained for each project;
- Each page of the notebooks/registers should be numbered;
- All entries should be dated, legible, clear, made in ink, and in a chronological and consistent manner (for instance, each new work day should begin on a new page);
- Leaving blank lines between entries should be avoided;
- Errors and deletions should be lined/marked, dated and initialled (and never erased out) in a manner that all entries remain readable and provide a quick visual account as to when and by whom the changes or errors have been corrected;
- Provide information to permit future verification of what was done by whom (appropriately initialled by the person making entry and verified by the supervisor);
- Any supporting materials or records should be properly catalogued and the reference to location(s) duly included in the notebook.
- *Photo of plots/fields vitiated due to unforeseen/unavoidable reasons- pest outbreak, flooding, animal grazing, on any other unforeseen calamity, etc. should be kept.*
- *It would be necessary to ensure that the data generator signs in the field note book at the end of each set of observations recorded. The PI/Co-PI must countersign it. HoDs may also countersign the note books periodically.*

4. Resources Required for Implementation

4.1 Infrastructure (Data input systems and Servers/Data Centre/back-up/ Disaster recovery, security systems), Responsibility (assigning roles)/Skills (to be imparted) should be in place

4.2 Computers at Institutes may be upgraded on priority as per needs..

4.3 Hand held devices may be made available to each scientist/technical officer for recording the data with geo-reference. This type of system requires some handy equipments such as tablets, scanner barcode reader which can perform in offline as

well as online mode. Though it costs a bit but the quick flow of accurate information may compensate the cost.

5. Operationalization of the Guidelines

To operationalize the guidelines, resources need to be allocated. Existing funds can easily be utilized for training of scientists at various levels. However, additional funds are required for purchase of Tablets and other accessories, software development, etc. A tentative cost is indicated below:

- **Hand-held devices/Tablets**: cost effective, robust hand-held devices/tablets with a built-in software for data recording. To reduce the number of hand-held devices/tablets, provisions of handling the same device/node by several researchers could be easily made using security features.
- **Barcode Reader and Printer:** Barcodes can be generated for each plot/experimental unit along with information such as experiment number, treatment, etc. To print barcodes on small labels, which can be pasted on individual plants/experimental unit (if required) or on a sign board using suitable media (such as glossy) so as to withstand weather conditions, barcode printers would be required. There is no need of a bar code scanner, in case tablets have the facility of scanning and reading the barcodes, otherwise, barcode readers would also be required.
- A mobile app is required to be developed for data entry/ or existing one could be customized.