International Water Management Institute
(IWMI)

Research Data Management Policy

And

Implementation Guideline

February, 2012

Applied to:	All research staff and projects
Administered by:	GIS, RS & Data Management (GRandD) Unit
First issued and approved:	December 2011

Current edition

Number:	1
Prepared/amended by:	Salman Siddiqui, Joanna Kane-Potaka, Sanjiv De Silva
Reviewed and cleared by Management Team:	December 2011
Approved by Board of Governors:	To be submitted

Summary

This document aims to outline the policy of the International Water Management Institute (IWMI) for research data management, and to provide guidelines to implement this policy at individual, project and institutional level. IWMI's research data management policy states that all research data acquired by researchers and projects are owned by IWMI. The following are some key points from the policy and implementation procedure:

- Projects must have a data management plan as part of the project work plan, containing a system for data collection, processing and archiving, and identifying the resources required for these activities.

- Each project will identify a **focal person** for data management of that project. This person will ensure that project data management guidelines, described in section 2.2 of this document, are followed in the project.

- All data must be archived in IWMI's central repository Water Data Portal (WDP) with standard metadata as early as possible after collection and processing with appropriate access rights.

- In order to maintain confidentiality of data in the central repository, different levels of access rights (individual, project, Institute or public) has been created. Project-level access rights for a dataset will be assigned based on the recommendation of the Project Leader (PL). On completion of the project, the PL may request closed access for a dataset for a maximum of one year to allow time for researchers to write research papers. After this period, all data will be accessible by public users. There may be exceptional cases to this, for which approval can be granted by the Deputy Director General (DDG)-Research.

- Regional offices will be provided a copy of relevant data for the respective regions, which will allow quicker access to researchers in those regions. They will also have a copy of metadata to search for the availability of data in offline mode. The offline metadata will be updated occasionally.

- Project leaders will be responsible for ensuring implementation of data management policy at project level while theme leaders will ensure that all projects under the themes are complying with the policy.

- The GIS, RS and Data Management (GRandD) unit will be responsible for developing standards, protocol and training to researchers and research assistants to use these standards.

- The GRandD unit will coordinate with projects for uploading data into the central repository and make them available to appropriate users.

- The annual performance evaluation should include research data management targets.

- As extra checks to ensure that data has been archived in WDP, there will be:
  - reporting on data management as part of standard project reports, including intermediate, annual and completion reports. The format of the document that should be used is given in Annex 1; and
  - data handover will be considered as part of the standard exit procedure for all IWMI staffs. This should follow the steps described in section 2.4 of this document.

1 Research Data Management Policy

1.1 Definitions

Research Data include primary or 'raw' data (recorded information, resulting from original observations and activities of a study), derived data (data that were originally 'raw' but were then converted to another form using some manual or automated process) and secondary data (acquired/collected from third-party sources). Forms of research data include information in laboratory notebooks, field notes and journals, survey questionnaires and photographs, and the data that describe them, GPS data, statistical package output files and satellite imagery. This document primarily refers to data in digital formats.

Research Data Management (RDM) involves the planning, development, implementation and administration of systems for the acquisition, storage and retrieval of data. It is a process that begins with the conception and design of a research project, and continues through data capture and analysis to publication, data archiving and sharing with the broader scientific community, as appropriate.

Metadata is a description of data and is required to discover, retrieve, understand and use the data. It is prepared in a way that can be used by anyone to find out, with reasonable effort, what data there is and whether it is of interest to them, how to get access to the data and what (if any) constraints exist using and distributing the data.

Data Originators are the designers and implementers of the research data collection. They are responsible for data quality and storage; they also maintain and update all metadata information related to their research data.

Data Analysts are the focal person for looking after data within each project**.** Each project will identify one person as data analyst for the project who will support data originators and other stakeholders to meet all elements of this policy. They are responsible for data organization, metadata preparation and submission of data to the institutional archive.

Data Focal Points are members of the data management core group and are responsible for uploading data in the archive and maintaining data beyond the scope of the project. They will provide assistance and look after data within each regional office**.**

Central Repository is the central storage facility maintained at IWMI HQ and refers to the Water Data Portal (WDP) which is an online database with a set of services to capture, store, index, preserve and redistribute IWMI's research data and outputs. It provides one stop access to all IWMI data in compliance with copyrights, intellectual property rights and data agreements.

Introduction

1.1.1 Background

The context in which data and information is managed has changed considerably over the last two decades. The quantity and diversity of data has increased to meet the growing demand for in-depth disciplinary and integrated multidisciplinary studies. Data generated using public funds is increasingly being recognized as public property. Donors are demanding a higher level of transparency in research methods and data collection. There is a growing demand for secondary data to support retrospective and prospective studies. Use of secondary data facilitates the integration of regular monitoring datasets, generated by government agencies, to be incorporated into scientific studies.

The mission of IWMI is to generate and share knowledge on land and water resources management in developing countries. The central focus of IWMI's knowledge management is to produce global public goods, and widely disseminate its data and research findings appropriately.

1.1.2 Why we need this policy

A survey was conducted among IWMI researchers during the Annual Research Meeting (ARM) in 2008 about research data management at the Institute. Most of the respondents supported that data acquired through projects should be centrally retained and accessible for future use. At the same time, they thought that the current organizational policy doesn't promote good data management, and the procedures are not adequate to ensure prevention of data loss when staff leave the Institute. At the institutional level, IWMI recognizes the need to maximize its investment on data acquisition and is able to extract information from data in the long term, to make previously unanticipated analyses and to compare it with future data. Donors have started tying public research funds to effective data management practices, e.g., AgWater Solutions Project (Bill and Melinda Gates Foundation). This policy is a response to the demands made by researchers and management to ensure that the most value is obtained from our research data.

1.1.3 Scope of the policy

The policy applies to all 'Research Data' at IWMI.

There are few policies at IWMI that in some way refer to 'data' (e.g., ownership in an IP policy, personnel policy). This policy attempts to focus on issues relating to research data. Consequently, the policy must be adhered to in conjunction with the other relevant policies of the Institute.

The policy would apply to any research project where IWMI is responsible for research data management. In the case of a research project where a partner is responsible for research data management, then the project contractual agreement must specify the principles that will guide the data management strategy. In the absence of such principles, then this policy will apply.

1.2 RDM policy goal

The goal of this Research Data Management policy is to maximize the benefit from research data by using it in present and future research and making it available as global public goods without violation of other's copyright and confidentiality.

1.3 RDM principles

This section describes the principles of research data management at IWMI. The guidelines to follow these principles at project level are described in section 2.2.

1.3.1 Ownership

Unless otherwise specified, all 'Research Data' generated or collected by IWMI research activities belong to the Institute. Ownership of 'Research Data' collected in collaborative projects is co-owned unless otherwise specified in the agreement document. Projects are custodian for data acquired under that project. Data originators and subsequent users are also custodian for the data. Data provided by third parties are subject to ownership of that party. IWMI will comply with the data policy of the respective agency and the agreement under which data are provided.

1.3.2 Data access and dissemination

IWMI's institutional objective is to produce maximum value from the data no matter who collected it.

- All research data (primary or secondary) will be submitted to the central repository immediately after they are collected and processed.
- During the project implementation period, project team members, authorized by the PL, will have access to data acquired by the project. During this time, other IWMI staff will be allowed access to data with the permission of the PL only.
- At the end of a project, all IWMI staff will have access to data from that project. Project leaders may request suspension of data release to the public domain for a maximum period of one year to allow time for researchers to write research papers. After that period, all data will be released to public users. If the 'data originators' or project leaders want to make an exception for a particular dataset, they will have to obtain written permission from the DDG-Research.
- Secondary data having copyright restriction from the sources will be accessible in strict compliance with the restriction of the respective sources.
- For any use of a dataset, the data users must properly acknowledge the 'data originators'.

1.3.3 Data archive

All Research Data generated from research activities will be stored in the long-term institutional archives. The availability of data in the archive will be published through the online system to IWMI users. Potential users will be able to retrieve data, subject to the restrictions stated in Principle 4.2.

1.3.4 Metadata

Metadata will be prepared according to ISO standards (e.g., 19115 and 19115-2 for spatial data) and assigned to all data so that potential users have sufficient information to decide whether it meets their requirements and whether they wish to use the data.

1.3.5 Data quality

Projects will implement explicit data validation procedures for all data generated in the project. Data collected from secondary sources will be screened to determine the originality of the dataset.

1.3.6 Efficiency

Projects will implement data management procedures that are designed to improve the efficiency of the research process. This will also help in using standard tools and sharing those among projects.

1.3.7 Ethics

Confidentiality of data on human subjects will be respected in accordance with national data protection acts (where they exist) and well-established international standards. The project will have measures to prevent the use of fraudulent data (see section 2.2.6).

2 Implementation Guideline

2.1 Implementation framework

A viable framework for managing research data that was developed and used by ICRAF (2008) is shown in Figure 1. This describes four key function categories: (i) project-level functions, (ii) institute-level functions, (iii) support functions, and (iv) audit/monitoring functions.

Project level

- All the research activities at IWMI are managed through projects, so research data must primarily be managed by respective projects and subsequently transferred to the institute level to provide institutional and public access.

- The project will recognize RDM as a planned activity and allocate resources accordingly. The RDM plan should be an integral part of the project work plan from the beginning and should be one that could be modified during the project period.

Institute level

- IWMI will recognize data management as a defined specialization in the center and assign/hire people to build on this specialization. This would be a research support skill for database design, data organization, data manipulation, metadata preparation and data archiving. These specialized staff will be part of the GRandD Unit and will provide services to projects in this area.

- An institutional-level data archive will be maintained centrally with a decentralized subset at regional offices through nodes. All the Research data will be stored in the central archive accessible through the online interface. To ensure rapid access at regional offices, a relevant subset will be stored in the respective offices. Periodically, the two systems will be synchronized. The regional offices will have a copy of complete metadata to explore data availability in offline mode. The metadatabase will be synchronized with the central metadatabase, periodically.
- Theme leaders will ensure that all projects under the theme are complying with the data management policies and archive their data in the institutional archive.

Support functions

- Data administrators from the GRandD Unit will coordinate with projects for uploading project data in the central archive and making it accessible to users.

- The GRandD Unit will develop standards and protocols, and provide training to researchers and research assistants to adopt the best practices in data management.

Audit functions

- The annual performance evaluation will include research data management targets, as stipulated in the data management plan.
- Projects will be selected for ‘auditing’ using the Research data management plan to evaluate data management standards within the project.

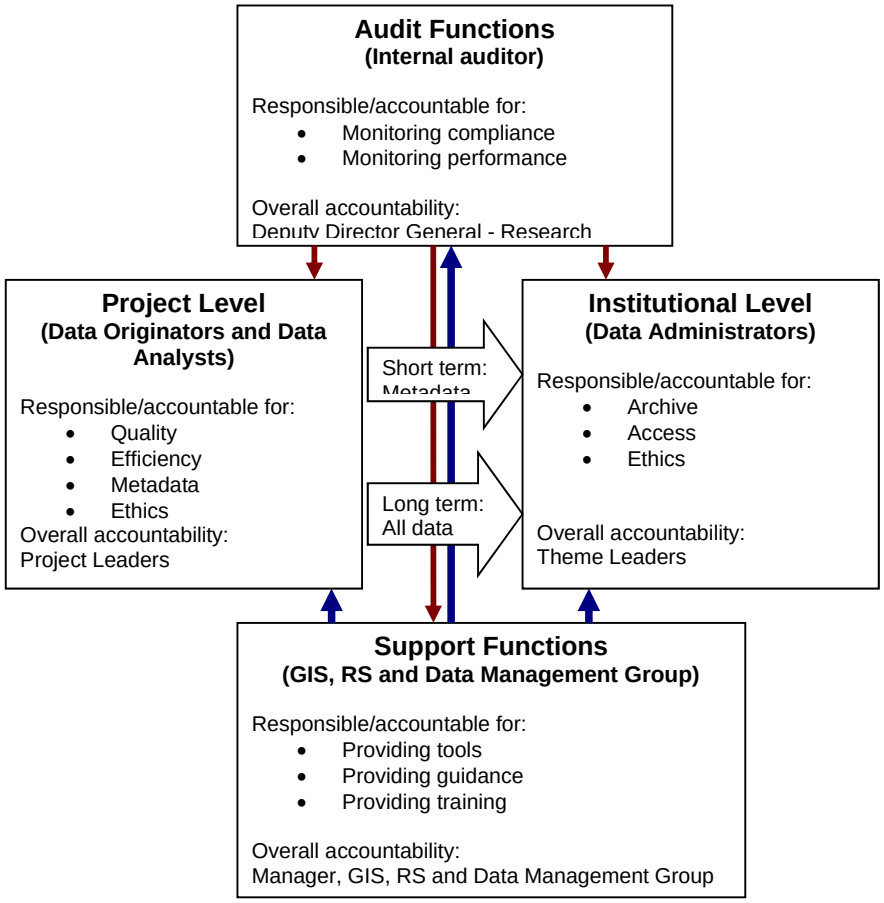

Figure1: The implementation framework (*Source:* ICRAF 2008)

2.2 Data management at project level

Project leaders are responsible for establishing an overall plan for the management of data under each project. The project data management plan will include the following:

- Specific activities with a time line including data collection, processing, analyzing, and archiving and budgeting for those activities.
- Identifying data management needs including resources and training for a project.
- Assign a Data Analyst to be the focal person for data management of the project. This could be the project leader or any person from the project team. This role will be recognized in their Indicators of Performance (IOPs).

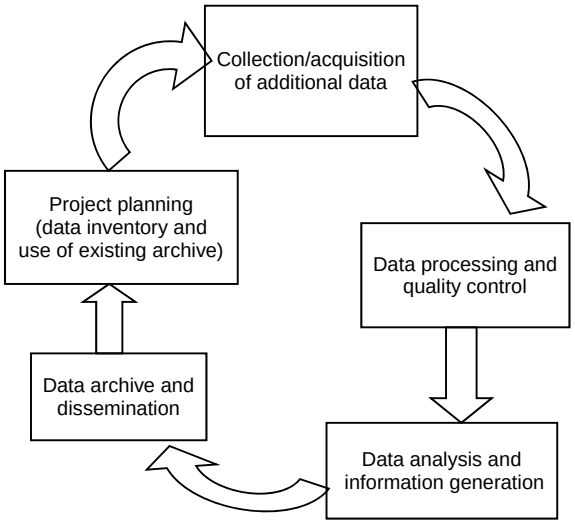

Figure 2: Project data management cycle

2.2.1 Data access and dissemination

Data collected/acquired by projects should be disseminated for wider use as early as possible to ensure the maximum value is obtained from the data.

- Sharing data during the project period: Projects are considered to be the custodian of all the data acquired under the project until that data are sent to the institutional archive. Project members are expected to ensure that the maximum value is obtained from the data by sharing it among team members and similar concurrent projects while maintaining good relationships through respecting expectations and agreements.
- Rights of use: No IWMI scientists have exclusive rights to the use of data under their custody. However, the data originator's expectation to be the first to publish using the data should be respected.
- Data sharing among partners: In partnership projects, data ownership and rights of use should be specified.
- Handling third-party data: Data collected from third-parties should be handled according to mutual agreement with the provider.
- Handling confidential data: Data that have potential risk of violating individual confidentiality should be handled in compliance with national laws.

2.2.2 Data archiving during and after projects

All research data are subject to archiving in the institutional repository of IWMI unless a project leader decides that there is no value to this and it is approved by the DDG-Research. The following issues will be addressed while archiving project data:

- Time of archiving: During the project period, a live archive will be maintained and updated by the responsible data analyst assigned to the project. At the end of the project, the complete datasets will be transferred to the central repository using the project closure data handover form shown in Annex 1. All data in the archive will be secured by assigning access rights suggested by the project leader or data originator.
- Access: Metadata for all data in the central repository (WDP) will be accessible to all IWMI staffs. In general, all original data will also be accessible to all IWMI staffs unless restricted by third-party copyright.

- Storage: All data will be stored in the central data archive unless there is a reason not to store it at IWMI headquarters. Regional subsets of data will be stored at regional offices to allow quicker access to data that is relevant to that region. Metadata will be stored in a central metadatabase and accessible online through the internet.

Please follow the steps below when submitting data into the WDP:

- Organize all the related data files (data file, code sheets) together.
- Collect all the relevant documentation together including the questionnaire, survey report and any other reference materials.
- Prepare metadata for the dataset as described in section 2.2.3 including the following minimum elements:
  - Dataset name, abstract, keywords, data source.
  - Methodology used for creation of the dataset, survey design, sampling method and size of the sample, and geographic and temporal coverage.
  - Information on data quality.
  - Access and use restriction, and contact person/organization.

2.2.3 Preparation of metadata

For future understanding, all archived data should include standardized metadata linked with each of the datasets together with supplementary documents that will be useful to understand the use of data.

The metadata should be prepared based on agreed standards and should explain the following:

- Identification: Dataset name, abstract, keywords and data source.
- Purpose: The purpose of acquiring/collecting the data.
- Process: The process of data collection including protocols (survey questionnaire), data collection plans, size of the sample and location referred to in the data.
- Source: includes contact information of data originators and data providers.
- Format: file format used for data storage. In case the file is stored in a proprietary format, the process for converting it to standard format (tools/software required).
- Time: Time period the data are referring to.
- Structure: Description of data structure (e.g., in case of tabular data, column definition).
- Organization: File structure (Name of the file with the description).
- Information on data quality.
- Access and use restriction, and contact person/organization.

Supplementary documents including project data collection plan, survey methodology, sampling procedure, data dictionary, etc., should be included in the metadata.

2.2.4 Data quality control

Data quality control (QC) procedures will be specified well before data are collected. This should be part of the research data management plan for the project. The following issues are important for data quality control:

- Data completeness: ensure all data collected according to plan are recorded and made available to the project data repository.
- Quality control: Action taken for correcting old data or deriving new data must be traceable.
- Data integrity: Avoid data duplication except for backup.

Project leader will prepare and include quality control procedures in the project work plan. Data originators will follow it and Data Analysts will verify that quality control measures stipulated in the RDM have been applied when data are accepted to the project repository.

Following are some guidelines for the quality control of secondary and primary data processing:

Secondary data are collected from various public and proprietary sources. In many cases these data require further processing (extraction, formatting) for use by the researchers. These data are of two categories: time series and non-time series. Quality control for secondary data includes:

- Converting data in a spreadsheet or a database.
- Use standard unit of measurements for time and other measures.
- Prepare data dictionary for column description, code value and unit of measurement.
- For location name, administrative unit name, etc., use standard names from authentic sources (e.g., name of district from national statistical division).
- In case of calculated fields, document the process (e.g., equation for interpolation of population).
- Avoid any text in the column that stores numerical data while using spreadsheet.

Primary data are collected through a field survey using a questionnaire to support in-depth research. To facilitate proper data management for survey data, the following issues should be considered:

- Design the questionnaire so that it is suitable for computerization
    - Use closed questions as much as possible.
    - Use skilled office coders for complex questions that can't be pre-coded.
    - Use a consistent coding scheme across the questionnaire.
    - Use a unique identifier for each survey unit and sub-units.
- Quality control during data entry
    - Check valid range in answers through automated validation.
    - Consistency check among related questions.

2.2.5 Efficiency in data processing

Data processing/analysis has to be fast and repeatable; preferably it should be reusable. The following issues are important for improving data management efficiency:

- Use of software: Software widely used at IWMI for particular types of work are preferable. Software should be interoperable with industry standards for the type of task. Data stored or manipulated should use interoperable formats.
- Data manipulation: Programmable methods should be used for data manipulation (e.g., formula in spreadsheets, script in access). In the case of a highly reusable process, it is preferred to develop custom tools and methods.

2.2.6 Ethical practices for handling data

All the persons involved in data handling at any stage are expected to follow ethical practices. They should respect the confidentiality and refrain from fraudulent use of data.

Confidentiality: When personal data are collected, the researcher should clearly explain to the concerned person about the purpose and use of that data.

Fraud: Researchers will not fabricate, ignore or change the original data that contradict reported findings.

2.3 Reporting on data acquisition and processing

Reporting on data should be part of standard project reports, including six monthly progress reports, annual report and a completion report. The project progress report should use the format provided in Annex 2 for reporting data management activities.

2.4 Data handover on staff exit

Data handover will be part of the standard exit procedure for all IWMI staff. The Human Resources (HR) Department and the Programme Office will coordinate together with theme leaders and project leaders to ensure that staff members do not leave IWMI with data (which belongs to the Institute), not having left a *useable* copy behind for further use. The following steps will be followed:

1. For each project the researcher is involved in, the project leader will assign a suitable person for handing over the project-related data. In a situation where the project leader leaves the Institute, the person who is taking over or the supervisor will assign someone to take over the data.
2. The researcher will organize the data files with related documentation (code sheet, questionnaire, description) and hand over one copy to the assigned person and one copy to GRandD Unit.
3. The researcher will complete the form in Annex 2 and submit it to HR for final clearance. A copy of the form will be given to the GRandD Unit for reference.
4. Such forms will be sent to the GRandD Unit through regional focal points from regional offices, when staff members in those offices leave the Institute.

References

CIFOR (Center for International Forestry Research). 2007. Data management policy, December, 2007.

ICARDA (International Center for Agricultural Research in the Dry Areas). 2005. Research information sharing and use policy.

ICRAF (World Agroforestry Centre). 2008. Research data management and archiving policy. World Agroforestry Centre, ILRI Policy Guidelines Series, June, 2008.

IFPRI (International Food Policy Research Institute). 2005. Scientific data standards and exchange: Capacity development. Final report, October, 2005.

IWMI (International Water Management Institute). 2004. Intellectual property policy, December, 2004.

Annex – 1: Format for reporting on data

Please note that any data collected under any project has to be made available to IWMI in a usable format and with appropriate documentation. Please contact Manger GRandD Unit if there is any query.

Please provide details of all data collected/acquired by the project in the format given below. If you don't have data on any of the types mentioned below, leave it blank. In case of data collected on any types not mentioned please add the type and use similar format (e.g. title, country, basin, time period, data source, use/distribution restriction, file name etc.) Please add any supplementary/additional comment you would like to mention on any data set. You will be contacted by Data Administrator in case any further clarification is required on any of the supplied data set.

Project Name:

Project Code:

Name of the Project Leader:

1. Secondary data (*List any data collected from secondary sources*)

1.1 Time series data on water and climate:

Title of the data set	Country	Basin	Number of station	Time period (From – To)	Data source	Restriction of distribution/use	Data files (Names of files) *

* Data files should contain station wise time series data, list of stations with station name, station id (if applicable), country, state, district and geographic position of stations (x,y,z)

Example: Daily stream flow, Monthly average rainfall

1.2 Statistical data on socio-economy, demography, agriculture, water, climate and environment

Title of the data set	Country	Basin	Number of location	Time period	Data source	Restriction of distribution/use	Data files (Names of files)*

*Data files should include actual data, list of locations including (country, state, districts…), field descriptions including name, unit of measurement, code lists (if contains coded value)

Example: District wise population by age group, State wise crop production and yield by crops

1.3 Base GIS/RS data on administrative boundary, transportation, hydrography, infrastructure, soil, land use, topography)

Layer name	Description	Geographic coverage	Format (vector, raster)	Feature type (polygon, line, point)	Data source	Source scale	Data provider (name address, e-mail)	Restriction of distribution/use	Data files (names of files)*

*Data files should include geometry file, projection file, description of field including name, unit of measurement, code list (if coded).

Example: District boundary of China, Detailed soil type

1.4 GIS data on socio economic, demography, water, agriculture, climate, land, environment and disaster)

Layer name	Description	Geographic coverage	Format (vector, raster)	Feature type (polygon, line, point)	Data source	Source scale	Data provider (name address, e-mail)	Restriction of distribution/use	Data files (names of files)*

*Data files should include geometry file, projection file, description of field including name, unit of measurement, code list (if coded).

Example: District wise population density map, Crop suitability map

1.5 Satellite image

Title	Geographic coverage	Type of satellite	Resolution	Time period	Data source	Restriction of distribution/use	File names

2. **Processed data**

List any manipulated data processed under your project (e.g. Land use map generated from satellite images, projected yield and production, interpolated rain fall etc..)

Data set title:
Description:
Purpose:
Geographic coverage (country, basin):
Resolution (in case of raster data):
Original data used:
Process description (process used for manipulation):
Access/use restriction:
Data file name:

3. Primary data

List any field survey data collected by the project include the following:

Title of the survey:
Objective of the survey:
Time period:
Location (country, area):
Method of data collection:
Number of samples:
File names (Data files, questionnaire, list of survey location, supplementary information*)

Supplementary information includes code sheets, field description including name, unit of measurement, reference to questionnaire section, code list (if coded)

Prepared by:

Name: **Signature:**

Approved by (project leader):

Name: **Signature:**

Supervisor:

Name: **Signature:**

Annex – 2: Staff exit data handover form

Name of the Researcher:

For each project, please complete the following section:

Project name (code):							
Project leader:							
Name			Signature:				
Data receiver:							
Name:			Signature:				
List of data transferred:							
Sl.	Data set title	Description of the data set	Source	Area coverage	Temporal coverage	Status	File name(s)*

*Data files should include actual data, list of locations including (country, state, districts…), field descriptions including name, unit of measurement, code lists (if contains coded value)

Signature of the researcher:

References

IWMI (2004) : Intellectual property policy, December 2004

ICRAF (2008) : Research Data Management and Archiving Policy, World Agro forestry Centre - ILRI Policy Guidelines Series, June, 2008

IFGRI (2005) : Scientific Data Standards and Exchange: Capacity Development, Final report, October, 2005

CIFOR (2007) : Data management policy, December 2007

ICARDA (2005) : Research information sharing and use policy