# Annexes

# CIP's Open Data & Data Management Guidelines and Procedures

**Version 1.0**

These nine annexes complement CIP's Open Data & Data Management Guidelines and Procedures. This document is intended to help researchers to meet their responsibilities with regards to research data quality, sharing and security.

**Correct citation:**

Research Informatics Unit. 2016. Annexes: CIP's Open Data & Data Management Guidelines and Procedures V1.0. International Potato Center (CIP), Lima, Peru. 46 p.

## Table of Contents

## Research Data Glossary

The list explains some of the terms used to describe CIP's Open Data and Data Management Guidelines and Procedures. Some terms are interchangeable with common usage, and some have specific meanings in particular contexts. The terms below are explained as they are used within the CIP Support for Open Access and Open Data.

**Archive.** A service to record, organize, and store (digital) items in optimal conditions, with standardized labelling to ensure their longevity and continued access. The service is based on application of metadata, archiving policies, records management, and digital preservation actions.

**CGSpace.** CGSpace is a repository of agricultural research outputs and results produced by different parts of CGIAR and partners. It indexes reports, articles, press releases, presentations, videos, policy briefs and more. CIP's CGSpace is at https://cgspace.cgiar.org/handle/10568/51671

**Curation.** Curation is the act of managing digital items held within an archive over the long term. It is an active process, implying action on the part of the curators so that items remain secure, discoverable and accessible. Digital curation involves maintaining, preserving and adding value to archived items throughout their lifecycle. The active management of digital research data reduces threats to their long-term research value and mitigates the risk of digital obsolescence.

**Data Access.** The procedures by which any individual or organization can freely acquire and use datasets collected or generated by foundation grantees or vendors with funding provided by the foundation. Data access generally involves activities such as cleaning, storage and retrieval of data

**Data Citation.** Data repositories standardizes the citation of datasets to make it easier for researchers to publish their data and get credit as well as recognition for their work. When you create a dataset in Dataverse, the citation is generated and presented automatically. Data citation has the following: the author(s), title, year, data repository (or distributor), version number, and a persistent identifier.

**Data Management Plan (DMP).** A DMP is a formal document that outlines what you will do with your data during and after a research project. Most researchers collect data with some form of plan in mind, but it's often inadequately documented and incompletely thought out. Many data management issues can be handled easily or avoided entirely by planning ahead.

**Data Repository.** An online storage tool for datasets that meets the following items such data must be accessible for a minimum of 5 years, data should be easily discoverable through conventional search mechanisms by an informed lay person (e.g. researchers and graduate students in the field), metadata on the dataset should be made available, data must be anonymized to protect individual personal identifiable information, open data platforms should honor any special ownership and access preferences as agreed between the foundation and

the data producer; data access may be limited to a specific audience or granted on a case by case basis.

**Database.** A database is a structured set of data, accessed via a database management system. There are various different types of database, providing different ways of structuring information. The most common type is a relational database, which expresses both the properties of, and the relationships between, different instances of particular objects. Databases are useful tools for supporting research, as they enable researchers to structure data and then rapidly query it in a consistent manner.

**Dataset.** An electronically stored collection of data and associated files. The data contained in a dataset may be from primary data collection (e.g. a survey) or secondary data generation via aggregation or synthesis. Datasets may contain one or more files and should include files that contain the data themselves; that document and explain the individual variables; and that explain on the collection or synthesis methodology. Some of the information describing the data may be contained in 'metadata' stored with the dataset.

**Dataverse.** Dataverse is a commonly-used open source data repository platform that facilitates the ability to publish, share, reference, extract and analyzes research data. It helps to make research data openly accessible. Each Dataverse contains studies or collections of studies, and each study contains metadata that describes the data plus the actual data and complementary files. The local installation of Dataverse from CIP is at https://data.cipotato.org

**Digital Object Identifier (DOI).** A DOI is a persistent identifier that is usually assigned to a digital item such as an article or a dataset in order that the item can be found and cited. DOIs can be incorporated into URLs so that users can always access the digital content, even if it has moved online location.

**Discoverable (or Findable).** Datasets are discoverable when reference links to the datasets are included in online directories (e.g. from repositories); a reference link to the dataset is provided in any publications or reports, or on the project/institution website; and/or returned when running a standard internet search. A common internet search engine should return a clear description of the data and a working link to the dataset or the repository where the data are housed.

**FAIR.** Research outputs such as articles and datasets are: Findable (the information products include descriptive, meaningful metadata), Accessible (the information products are available for the public to access immediately and with no restrictions, Interoperable (the information products are deposited in standards-compliant repositories), and Re-Usable (the information products should use open licenses).

**Metadata.** Information stored electronically with or as part of the dataset and should be provided along with the data whenever they are downloaded, accessed, or shared. This may include items such as year of data production, content, data dictionary, known data quality profile/issues, data completeness, other salient features of the data and dataset, and methodology used to collect/compile/create data.

**Research Data.** Research data and records are defined as the recorded information (regardless of the form or the media in which they may exist) necessary to support or validate a research project's observations, findings or outputs.

**Research Outputs.** Reports, publications, scientific presentations, policy briefs, working papers that present summary statistics, analysis and conclusions derived from primary or secondary data. Research outputs are distinct from datasets. Reporting on or sharing research results and outputs (e.g. summary statistics or tables) fulfills some of the objectives of the foundation's global access principles, but does not satisfy the requirement of data access.

## Annex 1. Research Protocols

A protocol is a written document details exactly how the research activity will be carried out. Every research activity should have a protocol. In the initial stages of an activity, a protocol is the plan. During the activity, any deviations from the plan should be recorded in detail in an updated version of the protocol. At the end of an activity, the final protocol should be an accurate reflection of the actual activity, with enough detail to enable another researcher to repeat the activity, using the same methods, to exactly the same standards, using only the protocol.

CIP has adopted the ISA-Tab structure for organizing experiments and data (Annex 6, Data Storage and Archiving). The three key entities are the investigation, the study and the assay levels. The research protocol will be used at the investigation level. Manuals, guidelines and operational procedures will be used at the study and the assay levels.

### 1.1 Content of the Research Protocol at investigation level

The protocol should contain the following sections:

- ✓ A Title that should be descriptive of the investigation, but concise
- ✓ An Abstract that should concisely summarize the elements of the protocol
- ✓ Abbreviations and Acronyms
- ✓ Version information: This is particularly important as a protocol should be a plan of how an activity should be conducted which is then updated to incorporate any deviations to reflect what actually happened. A table such as this can be included in the protocol to keep track of changes and updates made to the document.

**Table 1. Version control table**

| Version Number | Update Details |
|---|---|
| 1.0 | Original draft |
| 2.0 | Updated to include comments from reviewers (scientist/s: name/s, statistician/s: name/s) |
| 3.0 | Updated to remove location X, no longer safe to work in that area |
| 3.1 | Updated team members name and contact detailed due to change in project management |
| … | … |

- ✓ General location information
- ✓ Investigators - this should include names, institutions and area of expertise.
- ✓ Background and justification: This section of the protocol should specify:
    - What is the problem?
    - How will your research address the problem?

- What are the next steps expected to be after this activity?
- Who are or what is the target group and how was this group identified?
- Who will feel the impact of your research?

✓ A literature review

✓ Hypotheses – statements about what you expect to see from the results of your activity.

✓ Potential impact – who will benefit, how and by how much, will the benefits be sustainable,are there any negative effects to take into consideration?

✓ Activity objectives – these should be detailed, consistent and achievable through the activity.

✓ Methods – ensure this section has enough detail so that at the planning stage colleagues can review and comment on how the activity is to be carried out, and also to leave no doubts in other team members' minds as to what they should be doing whilst conducting the activity. This section generally contains the following:

- Type of research activity: experiment, survey, or observational
- A timeline for the activity
- Specific location information including how the locations are to be / were selected
- Study units – households, fields, farmers, again detailing how they are to be / were selected.
- Interventions: such as treatments or specific conditions set up by the research activity to observe their results.
- Inputs – this should include 'materials' required for the activity, these could be fertilizer, seeds, or translated questionnaires.
- Management – detail who is responsible for which part of the activity.
- Statistical design
- Data collection – which are the key response variables for the activity, how are they to be measured and by whom (farmer or technician?), measurement units, ensure that there are unique identifier variables (Annex 4, Documentation & Metadata).
- Data quality and assurance - detailing how data will be controlled for quality assurance at prior (quality based on a research protocol) and posterior data collection (consistency with established internal and external data standards). (Annex 5, Data Quality).
- Data management – detailing who will be responsible.

✓ Analysis, reporting and feedback – this should include a descriptions of the methods to be used to analyze the data, details on how the primary and secondary outcomes will be analyzed, statistical methods to be used, who is going to carry out the analysis. Include weight information for the key variables of interest; also include details of how the results will be fed back to those who participated in the activity.

✓ Implementation plan – including tasks, lists of partners and their roles, budget information.

✓ Dissemination of results and publication. The protocol should specify not only dissemination of results in the scientific media, but also to the community and/ or the participants, and consider dissemination to the policy makers where relevant.

✓ References

- ✓ Annexes and appendices

## 1.2 Procedures, manuals and guidelines to be used at study and the assay levels

CIP has developed the following list of procedures, manuals, guidelines that can be used at study and the assay level. Procedures, manuals and guidelines are publications of detailed methodologies to increase the reproducibility of research results. If you need to create a new protocol follow the instructions described above in point 1. You may wish to expand some sections and attach supplementary material.

**List of procedures, manuals and guidelines published at CIP**

- ✓ Burgos, G.; Muñoa, L.; Sosa, P.; Cayhualla, E; Carpio, R.; zum Felde, T. 2014. Procedures for chemical analysis of potato and sweetpotato samples at CIP's Quality and Nutrition Laboratory. Lima, Peru. International Potato Center (CIP), Global Program Genetics and Crop Improvement. ISBN 978-92-9060-444-0. 32p. http://dx.doi.org/10.4160/9789290604440
- ✓ De Haan, S.; Forbes, A.; Amoros, W.; Gastelo M.; Salas, E.; Hualla V.; De Mendiburu F.; Bonierbale M. 2014. Procedures for Standard Evaluation and Data Management of Advanced Potato Clones. Module 2: Healthy Tuber Yield Trials. International Cooperators Guide. Lima (Peru). International Potato Center. 44 p. ISBN 978-92-9060-448-8. http://cipotato.org/wp-content/uploads/2014/09/006184.pdf#sthash.pJOv1wIa.dpuf
- ✓ Forbes, G.; Pérez, W.; Andrade Piedra, J. 2014. Field assessment of resistance in potato to Phytophthora infestans. Lima (Peru). International Potato Center (CIP). 35 p. ISBN: 978-92-9060-440-2. http://cipotato.org/wp-content/uploads/2014/06/006154.pdf#sthash.F7YJaA6Z.dpuf
- ✓ Mihovilovich, E.; Carli, C.; De Mendiburu, F.; Hualla, V.; Bonierbale, M. 2014. Tuber bulking maturity assessment of elite and advanced potato clones protocol. Lima (Peru). International Potato Center. 43 p. Available on https://research.cip.cgiar.org/confluence/display/GDET4RT/Protocols.
- ✓ Montesdeoca, F., Panchi, N., Pallo, E., Yumisaca, F., Taipe, A., Mera, X., Espinoza, S., y Andrade-Piedra, J. 2012. Produzcamos nuestra semilla de papa de buena calidad - Guía para agricultoras y agricultores. Centro Internacional de la Papa (CIP), Instituto Nacional Autónomo de Investigaciones Agropecuarias (INIAP), Consorcio de Pequeños Productores de Papa (CONPAPA), Fundación McKnight. Quito, Ecuador. p. 82. http://cipotato.org/resources/publications/book/produzcamos-nuestra-semilla-de-papa-de-buena-calidad-guia-para-agricultoras-y-agricultores/#sthash.1m395clf.dpuf
- ✓ Pérez, W.; Forbes, G. 2004. Technical manual potato late blight. ISBN 978-92-9060-391-7. http://cipotato.org/resources/publications/manual/technical-manual-potato-late-blight/
- ✓ Pérez, W.; Forbes, G. 2011. Guía de identificación de plagas que afectan a la papa en la zona andina. Centro Internacional de la Papa (CIP). 48 págs. ISBN 978-92-9060-402-0. http://cipotato.org/resources/publications/book/guia-de-identificacion-de-plagas-que-afectan-a-la-papa-en-la-zona-andina/
- ✓ Porras, E.; Burgos, G.; Sosa, P.; zum Felde, T. 2014. Procedures for sampling and sample preparation of sweetpotato roots and potato tubers for mineral analysis. Lima, Peru.

International Potato Center (CIP), Global Program Genetics and Crop Improvement. ISBN 978-92-9060-445-7. 13p. http://dx.doi.org/10.4160//9789290604457

- ✓ Production Systems and the Environment. 2013. Protocol for Designing and Conducting Potato Field Experiments for Modeling Purposes International Potato Center. ISBN 978-92-9060-430-3. http://cipotato.org/wp-content/uploads/2014/06/006092.pdf#sthash.Sokp4oid.dpuf

**List of accredited Operational Procedures at CIP**

These documents can be used at the Assay level. If you need to create a new Operational Procedures please follow the instructions described from the OP001 (Operational Procedure template)

- ✓ Falcon, R.; Simon, R.; Rojas, E.; Hirahoka, D. QMS - Document control procedure - OP001
- ✓ Panta, A.; Silvestre, R.; Franco, N.E; Vollmer, R.; Zea, B. QMS - Handling of complaints and positive feedback - OP002
- ✓ Panta, A.; Silvestre, R.; Franco, N.E; Vollmer, R.; Zea, B. QMS - Non-conforming work, corrective and preventive actions - OP003
- ✓ Panta, A.; Silvestre, R.; Franco, N.E; Vollmer, R.; Zea, B. QMS - Training - OP004
- ✓ Panta, A.; Silvestre, R.; Franco, N.E; Vollmer, R.; Zea, B. QMS - Control of equipment - OP005
- ✓ Panta, A.; Silvestre, R.; Franco, N.E; Vollmer, R.; Zea, B.; Zamudio, T.; Barrientos, M.; Cardenas, J.; Ramirez, C.R. Procedure for checking the performance of micropipettes, Balances and ph-meters - OP006
- ✓ Panta, A.; Vollmer, R.; Zea, B.; Franco, N.E; Zamudio, T.QMS - Audit - OP007
- ✓ Falcon, R. QMS - CIP's Comprehensive ISO Procedure: Acquisition, Maintenance and Distribution of in-vitro plant material incorporating appropriate plant pathogen screening techniques - OP010
- ✓ Rojas, E. RIU - Accession identifiers - OP011
- ✓ Zea, B.; Ynga, A.; Ruiz, M. Genebank - Pathogen elimination of potato - OP017
- ✓ Zea, B.; Llanos, C. ; Ruiz, M. Genebank - Pathogen elimination of sweetpotato - OP018
- ✓ Panta, A.; Silvestre, R.; Franco, N.E; Ramirez, C.M.; Zamudio, T.; Cruzado, J.; Sanchez, J.; Mendoza,V. ; Rojas, H. M. ; Mallma, V. Genebank - In vitro conservation of potato - OP025
- ✓ Panta, A.; Silvestre, R.; Franco, N.E; Ramirez, C.M.; Zamudio, T.; Mallma, V.; Rojas, H. M. ; Santa Maria, A.; Robles, O.; Murga, A.; Loayza, J. Genebank - In vitro conservation of sweetpotato - OP026
- ✓ Falcon, R. ; Grande, E. GADU - Acquisition of genetic resources - OP054
- ✓ Panta, A.; Silvestre, R.; Zea, B.; Santa Maria, A. ; Sanchez, J. Genebank - Introduction of potato to in vitro culture - OP055

- ✓ Franco, N.E; Vollmer, R.; .Zea, B.; Panta, A.; Barrientos, M.; Cardenas, J. Genebank - In vitro multiplication of potato - OP056
- ✓ Panta, A.; Silvestre, R.; Santa Maria, A.; Sanchez,J.; Zamudio,T.; Zea, B. Genebank - Introduction of sweetpotato to in vitro culture - OP058
- ✓ Franco, N.E; Vollmer, R.; Panta, A.; Barrientos,M.; Cardenas, J.; Zea, B. Genebank - In vitro multiplication of sweetpotato - OP059
- ✓ Vollmer, R.; Panta, A.; Franco, N.E; Barrientos,M.; Cardenas, J. Genebank - Distribution of in vitro material - OP068
- ✓ Zea, B. ; Llanos, C; Valverde, M.A; Ynga, A.; Fuentes, S.; Franco, N. E.; Monteblanco, T.; Panta, A.; Espinoza, I.; Ramirez, C.M. Genebank - Procedure for checking the performances of autoclaves and isotherms equipment - OP071
- ✓ Falcon, R.; Grande, E. GADU - Distribution of CIP genetic resources - OP072
- ✓ Panta, A.; Zea, B.; Cruzado, J.; Mendoza, V.; Robles, O.; Rojas, H. M.; Santa Maria, A.; Silvestre, R.; Zamudio, T.; Mallma, V.; Franco, N.E.; Ramirez, C.M.; Vollmer, R.; Barrientos, M.; Cardenas, J. Genebank - Preparation of Culture Media - OP074
- ✓ Genebank - Rules to calculate the CIP Distribution Status formula - OP085
- ✓ Del Villar, R. (CIP);Orue, R.l (CIP); Cordova, R. (CIP) ITU - Backup Management for databases supporting ISO accredited processes - OP086
- ✓ Ferreyra, E.; Ganoza, X.; Ramos, S.; Flores,S.; Tintaya, T.; Tinco, R. Logistics - Purchasing of services and supplies procedure - OP088
- ✓ A.Panta; R.Silvestre; N.E.Franco;B.Zea; C.Llanos; A.Santa Maria;A.Ynga; V.Mallma; O.Robles; J.Sanchez Environment Monitoring (screen houses, green houses or culture rooms) - OP091
- ✓ Ferreyra, E.; Cordova, S.; Alarcon, W.; Franco, M.; Pelaez, P.; Blanco, D. Logistics - Maintenance Procedure - OP093
- ✓ Franco, N.E.; Vollmer,R. Logistics - Subcontracting of Tests and Calibration - OP095
- ✓ Vollmer, R. Service to the customer - OP096

## Annex 2. Data Management Plan (DMP)

A data management plan is a formal document that outlines what you will do with your data during and after a research project. Most researchers collect data with some form of plan in mind, but it's often inadequately documented and incompletely thought out. Many data management issues can be handled easily or avoided entirely by planning ahead. With the right process and framework it doesn't take too long and can pay-off enormously in the long run.

### 2.1 Who requires a plan?

In November 2013, all 15 members of the CGIAR Consortium unanimously endorsed the Open Access and Data Management Policy[2] issued that Data Management Plans should be prepared in order to ensure implementation of the Policy. Such Plans shall, in particular, outline a strategy for maximizing opportunities to make information products Open Access.

### 2.2 Template outline

By getting to know your research and data, we can match your specific needs with data management best practices in your field to develop a data management plan that works for you. If you do this work at the beginning of your research process, you will have a far easier time following-through and complying with funding agency and publisher requirements. You should get familiar with the funding agency open access and open data policy. You should aim to comply with whichever is more stringent. We've compiled a list of DMP details from different funding agencies and plans specifically for how Data Management Plans are to be implemented. If a funding agency requires a data management plan, follow their requirements. Projects that are not required by a funding agency to follow a particular format should consider incorporating the elements listed here in their data management plans. Consider the following:

#### 2.2.1 Expected Data Type

Describe the type of data (e.g. digital, non-digital) and how they will be generated (lab work, field work, surveys, etc.). Are these primary or metadata? Data types could include text, spreadsheets, images, 3D models, software, audio files, video files, reports, surveys, etc. Consider the following:

- ✓ What data will be generated in the research?
- ✓ What data types will you be creating or capturing?
- ✓ How will you capture or create the data?
- ✓ If you will be using existing data, state this and include how you will obtain it.
- ✓ What is the relationship between the data you are collecting and any existing data?
- ✓ How will the data be processed?
- ✓ What quality assurance & quality control measures will you employ?

#### 2.2.2 Data Format

For scientific data to be readily accessible and usable it is critical to use an appropriate community-recognized standard and machine readable formats when they exist. The data should

preferentially be stored in recognized public databases appropriate for the type of research conducted. Regardless of the format used (notebook, samples, images, spreadsheet, etc.), that data set should contain enough information to allow independent investigators to understand, validate, and use the data. Describe the format of your data and how it will be "documented." Think about what information is needed for the data to be read and interpreted in the future. What would someone else need to be able to use these files? Consider the following:

- ✓ Which file formats will you use for your data, and why?
- ✓ What data will be preserved for the long-term?
- ✓ What transformations (to more shareable formats) will be necessary to prepare data for preservation / data sharing?
- ✓ What metadata/ documentation will be submitted alongside the data or created on deposit/ transformation in order to make the data reusable?
- ✓ What contextual details (metadata) are needed to make the data you capture or collect meaningful?
- ✓ How will you create or capture these details?
- ✓ What form will the metadata describing/documenting your data take?
- ✓ Which metadata standards will you use and why have you chosen them? (e.g. accepted domain-local standards, widespread usage)

### 2.2.3 Data Storage and Preservation

Scientific data should be stored in a safe environment with adequate measures taken for its long-term preservation. When applying for research funding, applicants should describe plans for storing and preserving their data during and after the project and specify the data repositories, if they exist. They should outline strategies, tools, and contingency plans that will be used to avoid data loss, degradation, or damage.

- ✓ Will you share data via a repository, handle requests directly or use another mechanism?
- ✓ If your method of sharing is with an archive, which archive/repository/database have you identified as a place to deposit data?
- ✓ What procedures does your intended long-term data storage facility have in place for preservation and backup?
- ✓ What is the long-term strategy for maintaining, curating and archiving the data?

### 2.2.4 Data Sharing and Public Access

Describe your data access and sharing procedures during and after the grant. Provide any restrictions such as copyright, confidentiality, patent, appropriate credit, disclaimers, or conditions for use of the data by other parties. DMPs should clearly articulate any justifiable limitations on project data sharing due to confidentiality, privacy, proprietary interests, business confidential information, and intellectual property rights and avoid significant negative impact on intellectual

property rights, innovation, and competitiveness. Any restrictions on data sharing, such as a delay of disclosing proprietary data, should be presented. Consider the following:

- ✓ Will any permission restrictions need to be placed on the data?
- ✓ With whom will you share the data, and under what conditions?
- ✓ Will a data sharing agreement be required?
- ✓ Have you gained consent for data preservation and sharing?
- ✓ Are there ethical and privacy issues? If so, how will these be resolved?
- ✓ How long will the original data collector/creator/principal investigator retain the right to use the data before opening it up to wider use?
- ✓ Explain details of any embargo periods for political/commercial/patent reasons?
- ✓ When will you make the data available?

### 2.2.5 Roles and Responsibilities

Who will ensure DMP implementation? This is particularly important for multi-investigator and multi- institutional projects. Provide a contingency plan in case key personnel leave the project. Also, what resources will be needed for the DMP? If funds are needed, have they been added to the budget request and budget narrative? Projects must budget sufficient resources to develop and implement the proposed DMP. Consider the following:

- ✓ Outline the staff/organizational roles and responsibilities for implementing this data management plan.
- ✓ How will responsibilities be split across partner sites in collaborative research projects?
- ✓ What process is in place for transferring responsibility for the data?
- ✓ Who will have responsibility over time for decisions about the data once the original personnel are no longer available?
- ✓ What costs if any will your selected sharing method charge?

### 2.2.6 Monitoring and Reporting

Successful projects should monitor the implementation of the DMP throughout the life of the project and after, as appropriate. Implementation of the DMP should be a component of annual and final reports and include progress in data sharing (publications, database, software, etc.). The final report should also describe the data that was produced during the award period and the components that will be stored and preserved (including the expected duration) after the award ends.

- ✓ Acknowledge that your project and DMP will be monitored. Who will be responsible for reviewing and revising this data management plan?

### 2.3 Data Management Plan Template:

This DMP contain a CIP template that outlines what you will do with your data during and after a research project. Specific templates for main donors are documented in “Donor Open Access Information” at //cipotato.org/open-access/.

**Data Management Plan Template.**

| | |
|---|---|
| **Project Title:** | |
| **Project Lead Center:** | |
| **Project Investigator:** | |
| **Individual responsible for Data Management:** | |
| **Donor** | |
| **Agreement Id or cost Center** | |
| **BUS** | |

## I. SUMMARY:

Describe your approach to Open Access and Data Management and any arrangements already in place to assist implementing the CIP Open Data and Data Management Policy:

## II. DATA AND DATABASES:

1. Describe the nature and scope of the data that will be generated under the project:

2. Describe your anticipated manner of storing, managing and making the project data and metadata available:

3. Describe who will have access to project data, and under what conditions *(if any)*:

4. Set out your plan for automated metadata capture and checking, the standards you will use and the steps you will take to ensure interoperability across information technology systems:

5. What is your anticipated timing for making project data and metadata available?

6. Explain whether any of the project data will not be made publicly available and why *(e.g., ownership or access to pre-existing data; license rights to project data; personal privacy concerns; competitive advantages; data sovereignty)*:

7. Describe the anticipated benefits and uses that could be achieved by making the project data available in the manner you describe above:

8. Describe any potential disadvantages of making the project data available in the manner you describe above:

**III. VIDEO, AUDIO AND IMAGES:**

Set out any repositories you will use for storing and sharing these information products:

**IV. COMPUTER SOFTWARE:**

Set out any repositories you will use for storing and sharing your software codes:

**V. RESOURCES AND BUDGETING:**

1. Describe the anticipated total costs involved with making data widely available *(if any)*:

2. What other additional resources or support will you require to ensure this Data Management Plan is delivered?

## Annex 3. Budgeting and Planning Template

This is an example budget to help determine costs and needs to make research outputs open.

### Annex 3. Budgeting and Planning Template

**I Communications & Marketing**

Consult with CPAD.

**II Open Access** **0**

| | No. of Publications | No. of Publications | Total Budget |
|---|---|---|---|
| **II.1 Peer reviewed articles/Publications** | | | |
| Article Processing Fees [1] | | | 0 |

| | Staff Hours from RIU team | | Total Budget |
|---|---|---|---|
| **II.2 Data** [2] | | | |
| Data Management Plan: support in the development of your project DMP | | | 0 |
| Documentation and metadata. | | | 0 |
| Data Quality | | | 0 |
| Data Privacy & Confidentiality | | | 0 |
| Other (please specify) | | | 0 |
| Total | | | **0** |

(1) Refer to Journal Data Base for Open Access *insert link*
(2) Refer to CIP Open Data & Data Management Policy and Guidelines *insert link*

## Annex 4. Documentation and Metadata

### 4.1 Introduction

At the basic level metadata is information that enables the user to fully understand a dataset or a resource. It is often described as “data about data”. The metadata promotes open access and data sharing. Your metadata should address these questions at sufficient level of detail.

#### 4.1.1 Investigation, Study and Assay level

Some of the questions are study level questions so we would have one level of metadata to describe the study – we will refer to this as the study catalogue. Think in terms of who, what, when, and where. For example:

- ✓ What is the study about?
- ✓ When was the study carried out? Give start and end dates if known.
- ✓ Where was it carried out? Indicate the location.
- ✓ Who carried out the study? Who were the individuals or organizations involved?

#### 4.1.2 Data Level

Other questions such as what does each column represent, are data level questions – we will refer to this level as the data dictionary. For each variable in your dataset you should indicate.

- ✓ The data type,
- ✓ Include a variable label to briefly describe it,
- ✓ If it is a coded variable then value labels will be needed so users can interpret the data correctly
- ✓ Have you used any missing value codes and what are they.

### 4.2 Core Metadata Schema

There have been many metadata initiatives that have attempted to standardize metadata content and format. Two of the most well-known are the "Data Documentation Initiative" (http://www.ddialliance.org/) and the "Dublin Core® Metadata Initiative" (http://dublincore.org/).

The CGIAR and CIP have adopted the original 15 Dublin Core elements. Additionally, CIP has added 3 additional elements. The qualifier ‘dc’ comes from Dublin Core, ‘cg’ refers to “CG Core,” and the ‘cc’ comes from CIP Core. The following list provides an overview of the elements, their status, and brief description of their usage. The following list is mandatory for the documentation at Investigation levels.

| Element | Qualifier | Definition |
|---|---|---|
| Title | dc.title | Official or unofficial title of the document, data set, image, etc.<br>The example includes type the experiment, year (YYYY), month (MM), location. Eg. PTYL200205_CIPHQ |

| | | |
|---|---|---|
| Creator | dc.creator | Creators of the item typically a person. Could be an organization in case of corporate authors (e.g. Center reports) |
| Subject | dc.subject | Subject matter of the research, technologies tested, etc. |
| | cg.subject.agrovoc | AGROVOC subject matter or research area |
| | cc.subject.crop | Cultivated plants or agricultural produce, such as tuber, roots, grain, vegetables, or fruit, considered as a group, allows the registration of crop used in the experiment or bioassay in laboratory, greenhouse or field. (E.g. Sweetpotato, potato, ullucus, etc.) |
| Description | dc.description.abstract | Abstract or longer description of the item |
| Publisher | dc.publisher | Entity responsible for publication, distribution, or imprint |
| Contributor | dc.contributor | Person, organization, or service making contributions to resource content; ; CGIAR affiliation |
| | cg.contributor.center | Research Centers and offices with which creator(s) are affiliated |
| | cg.contributor.crp | CGIAR Research Program with which the research is affiliated |
| | cg.contributor.funder | Funder, funding agency or sponsor |
| | cg.contributor.partnerId | Partners, funding agencies, other CGIAR centers |
| | cg.contributor.project | Name of project with which the research is affiliated |
| Date | dc.date | Publication or creation date: YYYY-MM-DD (confirm to ISO 8601) |
| | cg:date.embargo-end-date | Used when an item has an embargo by publisher (ex: 6 or 12-month embargo): |
| Type | dc.type | Nature or genre of item/content; e.g., poster, data set |
| Format | dc.format | File format of item e.g.: PDF; jpg etc. |
| Identifier | dc.identifier | Unambiguous reference to resource such as doi, uri |
| Source | dc.source | Journal/conference title; vol., no. (year) |
| Language | dc.language | Language of the item - ISO 639-1 (alpha-2) or ISO 639-2 (alpha-3) |
| Relation | dc.relation | Supplemental files |
| Coverage | dc.coverage | Geospatial coordinates, countries, regions, sub-regions, chronological period. |
| | cg.coverage.region | Supra-national areas (above country level) related to the item being described |
| | cg:coverage.country | Country/countries related to the data which was collected in resource (ISO 3166) |
| | cg:coverage.admin1 | First sub-national administrative division geography name (Department) |
| | cg:coverage.admin2 | Second sub-national administrative division geography name (Province) |
| | cg:coverage.admin3 | Third sub-national administrative division geography name (District) |
| | cc.locality | Community or population center |
| | dc.coverage.latitude | Geo-spatial location: latitude, Example: -10.76822 |
| | dc.coverage.longitude | Geo-spatial location: longitude, Example: -75.80617 |
| | cc.coverage.elevation | Geo-spatial location: altitude: 2918 msnm |
| | dc.coverage.start-date | Chronological period: start date of activity described in resource |
| | dc.coverage.end-date | Chronological period: end date of activity described in resource |

| | | |
|---|---|---|
| Rights | dc.rights | Rights, licensing, permission statement |

### 4.3 Core data dictionary

The phenotype data is the observable characteristic of an organism, such as its morphology, development, biochemical, physiological characteristic, phenology, agronomic characteristic and behavior. The genotype data is the genotypic information obtained by molecular technologies such as SSR, SNP, AFLP and others. Creating a data dictionary is vital to the standardization of variables, and the interoperability of data with different system.

#### 4.3.1 Phenotypic data dictionary

Crop Ontology is pursuing the development of standards for phenotyping data, in collaboration with other interested groups. Crop Ontology is intended to facilitate access to the data held within and/or across databases, in combination with a Trait Dictionaries for breeders' field books. Crop Ontology helps to harmonize data capture and manipulation through ontology-driven queries.

The following list provides an overview of the elements of an example of a Crop Trait Dictionary (TDv5) (See http://www.cropontology.org/). The following list is mandatory for the documentation at Study and Assay level in crop breeding.

| **Column** | **Description** |
|---|---|
| **Variable ID** | Unique identifier for the trait. If left blank, the upload system will automatically generate a trait ID. If a given trait is related to more than 1 variable, the trait ID must be specified and must be identical for the variables. |
| Variable name | Name of the variable following the convention <trait abbreviation>_<method abbreviation>_<scale abbreviation>. Variable name must be unique. |
| Variable synonyms | Other names, if any, given to this variable |
| Context of use | Indication of how trait is routinely used. If several "contexts of use", separate with "," |
| Growth stage | Growth stage at which measurement is made. Follow standards. If variable used in time series, leave blank |
| Variable status | Status of the variable. Possible entries are 'recommended', 'standard for <institution or community>', 'obsolete', 'legacy' |
| Variable Xref | Cross reference of the variable term to a term from an external ontology or to a database of a major system. |
| Institution | Name of institution submitting the variable |
| Scientist | Name of scientist submitting the variable. |
| Date | Date of submission of the variable. |
| Language | 2 letter ISO code for the language of submission of the variable. |
| Crop | Name of the crop for which the variable is recorded |
| **Trait ID** | Unique identifier for the trait. If left blank, the upload system will automatically generate a trait ID. If you want to create more than 1 variable for a given trait, the trait ID must be specified and must be identical for the variables. |
| Trait | Trait name (property) |
| Trait class | General class to which trait belongs. Consensus trait classes are 'morphological trait', 'phenological trait', 'agronomical trait', 'physiological trait', 'abiotic stress trait', 'biotic stress trait', 'biochemical trait', 'quality traits trait' and 'fertility trait' |

| | |
|---|---|
| Trait description | Textual description of trait. |
| Trait synonyms | Full text synonyms, if any, of the trait. If several synonyms, separate with commas. |
| Main trait abbreviation | Main abbreviation of the trait name. It is mandatory and has to be unique within a crop TD. By convention, this abbreviation must not start with a digit, must have no space. |
| Alternative trait abbreviations | Other frequent abbreviations of the trait, if any. These abbreviations do not have to follow a convention. If several alternative abbreviations, separate with commas. |
| Entity | A trait can be decomposed as "Trait" = "Entity" + "Attribute", the entity is the part of the plant that the trait refers to e.g., for "grain colour", entity = "grain" |
| Attribute | A trait can be decomposed as "Trait" = "Entity" + "Attribute", the attribute is the observed feature (or characteristic) of the entity e.g., for "grain colour", attribute = "colour" |
| Trait status | Status of the trait. Possible entries are 'recommended', 'standard for <institution or community>', 'obsolete', 'legacy' |
| Trait Xref | Cross reference of the trait to an external ontology or database term e.g., Xref to a trait ontology (TO) term |
| **Method ID** | Unique identifier of the method. If left blank, the upload system will automatically generate a method ID. |
| Method | (Short) name of the method |
| Method class | Class of the method. Entries can be "Measurement", "Counting", "Estimation", "Computation" |
| Method description | Textual and generic description of the method. |
| Formula | For computational methods i.e., when the method consists in assessing the trait by computing measurements, write the generic formula used for the calculation |
| Method reference | Bibliographical reference describing the method. |
| **Scale ID** | Unique identifier of the scale. If left blank, the upload system will automatically generate a scale ID. |
| Scale name | Name of the scale |
| Scale class | Class of the scale, entries can be "Numerical", "Nominal", "Ordinal", "Text", "Code", "Time", "Duration" |
| Decimal places | For numerical, number of decimal places to be reported |
| Lower limit | Minimum value (used for data capture control) for numerical and date scales |
| Upper limit | Maximum value (used for field data capture control). |
| Scale Xref | Cross reference to the scale, for example to a unit ontology such as UO or to a unit of an external major database |
| Category 1 | If the scale is categorical, entry must follow the convention: '<label of the category> = <meaning of the category>' or '<label of the category> = <abbreviation of the category> = <meaning of the category>' |
| Category 2 | If the scale is categorical, entry must follow the convention: '<label of the category> = <meaning of the category>' or '<label of the category> = <abbreviation of the category> = <meaning of the category>' |
| Category n | If the scale is categorical, class value and meaning of the n-th category. It possible to create as many category columns as necessary, as long as they are called "Category <number>" |

#### 4.3.2 Molecular data dictionary

CIP is developing formats to collect Molecular Data based on templates MIQAS_TAB. The MIQAS_TAB format is designed to describe the results from a QTL or association study according to MIQAS minimal requirements (http://miqas.sourceforge.net/specification/MIQAS_TAB/MIQAS_TAB_specification.html).

#### 4.3.3 New data dictionary

If you need to create a new Data Dictionary for other areas (social sciences for example) follow these instructions:

- ✓ Create an initial Data Dictionary (list of variables) in English.
- ✓ Verify if there are duplicate and redundant variables with a consultation of the team researchers.
- ✓ Complete the documentation for the selected variables (units, scale or method) in accordance with the research protocols.

### 4.4 Link to other dictionaries

The Thesaurus and Glossary are online vocabulary tools of agricultural terms in English and Spanish. It contains over 110,000 terms, including 51,926 cross references glossaries of definitions. http://agclass.nal.usda.gov/

AGROVOC Multilingual agricultural thesaurus is a controlled vocabulary covering all areas of interest of the Food and Agriculture Organization (FAO) of the United Nations. It is published by FAO and edited by a community of experts. AGROVOC consists of over 32,000 concepts available in 23 languages: Arabic, Chinese, Czech, English, French, German, Hindi, Hungarian, Italian, Japanese, Korean, Lao, Malay, Persian, Polish, Portuguese, Russian, Slovak, Spanish, Telugu, Thai, Turkish and Ukrainian. http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus.

## Annex 5. Data Quality

### 5.1 Data quality assurance

Quality assurance of data has at least two dimensions: prior and posterior.

- ✓ **Prior data quality** is defined as quality based on a research protocol.
- ✓ **Posterior quality** is defined as quality based on consistency with established internal and external data standards -- basically, anything one can do after the data have been collected.

Within a new project or activity it is most effective to assure quality of data in a preventive way; one such model is to establish a quality management system. This consists basically in three steps: (a) say what you do, (b) do what you say, and (c) have someone else check it. In a typical research setting, the first task is usually the responsibility of the scientist, the second corresponds to the technical staff, and the third is carried out by the data analyst or statistician. The more formal and documented the whole process is, the more checks are built in, and the expected quality is higher. Similarly, the more automated the process is - using well-maintained equipment - the more consistent the resulting data. Therefore, a first principle is to standardize and automate as much as possible and is reasonable.

#### 5.1.1 When to carry out data checks

Data quality is important because other researchers or experts can use the data in new analysis, meta-analysis, or different applications of modelling or statistical tools, leading to new insights for the future and so the data have a longer-term value as a resource to the research.

Data quality checking is the process of reviewing the data to discover inconsistencies and other anomalies and performing data cleaning activities to improve data quality. Quality control at different stages is part of setting up systems for Quality Assurance.

A dataset is almost never 100% clean, but the aim is to produce datasets that are as error-free as possible. If you have obvious mistakes in your data you may find you have difficulties justifying your results and conclusions. On the other hand, if you can produce proof that you have systematically checked your data and have eliminated as many errors as possible, then your results will have greater credibility. You should be aware that errors in data can occur at any stage of the data cycle:

- ✓ During the design of your data collection instruments and procedures. There is no actual data at this stage but you should be aware at this stage of the data you wish to collect and have in place mechanisms for minimizing errors later;
- ✓ During data collection – e.g. the measuring instrument was not properly calibrated, or it was read incorrectly, or the value was recorded wrongly.
- ✓ During data entry – data entry errors are common, for example it is very easy to hit the wrong key or put the decimal point in the wrong place. You can use double data entry (DDE): the data are entered twice, into separate files and by different individuals. The two data files generated from this process are then compared, and any discrepancies are

checked against the original data collection sheets or questionnaires. Additionally, you can define the range of values for the variables and data dictionaries/ontologies for the standardization of variables (Annex. 4, Documentation & Metadata).

- ✓ During data manipulation and analysis – e.g. values may be truncated when transferring between software; you may make a mistake in a calculation, etc.

RIU has developed the HiDAP Tool (High-throughput Data Analysis Platform) to help with the standardization of field trial datasets (collection, quality control, analysis and reporting). The tool can be used at https://apps.cipotato.org/shiny/projects/hidap

#### 5.1.2 Dealing with Outliers

Once the data are entered you can compare across records looking for extreme values and, in the case of categorical data, looking for values outside the range of possible categories. We would recommend running simple frequency tables for all categorical variables. Boxplots are useful tools to help visualizing and assessing the spread of the data and whether it clusters in the regions that would be expected.

#### 5.1.3 Transfer of data from field to base

As part of your data management plan and your data quality control, you should document how you intend transferring your data from the field to the office. This might involve the transport of paper questionnaires or collection sheets, or it might involve the method you intend to use to transfer data from hand-held devices. This includes the frequency of transfer and any systems you have in place for ensuring safe delivery.

#### 5.1.4 Electronic transfer and backups

When data are entered in the field you will need a method of transferring and checking the data. As files are transferred you need to ensure they arrive uncorrupted – there should be mechanisms for checking the file(s) on arrival – don't delete the file(s) from the hand-held device until successful transfer has been confirmed. Backing up your data is an important aspect of data management and data quality control. (Annex 6, Data Storage and Archiving)

#### 5.1.5 Audit trail

We recommend keeping an Audit Trail, i.e. a document detailing the checks run on the data and any corrections/changes made. Decisions made on how you dealt with any outliers should be included in this same document. If you have used double data entry then the results from the data comparisons can also be included in the Audit Trail.

#### 5.1.6 Units of measurements

A common problem when collecting data is to do with inconsistencies in the unit of measurement. We recommend to document adequately in the data dictionary as described in Annex 4, Documentation & Metadata.

#### 5.1.7 Versions of data

During the lifetime of a research project there are likely to be different versions of the data as errors are identified and corrected. Some researchers like to keep different versions particularly if they have started analysis before all the errors are identified. If you are going to keep previous versions you should develop a naming convention for your data files together with a log outlining

the differences between the versions. You naming convention might just be to include the version number in the filename or you might prefer to include the date in the filename. Either way the method you use should be documented and agreed by all parties involved in the project should be aware of which version is the latest or definitive version.

## 5.2. Guidance for handling 'special' types of data

### 5.2.1 Free Text Data

Free text data is text data that cannot be coded into categories, such as respondent name, or data resulting from open ended questions such as 'give your opinion…', 'do you have any further comments…', or enumerator comments. Free text should be entered exactly as seen in the original source; no text should be paraphrased, however non-ASCII text characters (e.g. accents, Greek letters etc.) should not be entered.

### 5.2.2 Missing Data

Where possible the study should be designed to allow for as much information to be collected as possible. Allow for options such as 'other, please specify', and 'NA', responses to be specified as well as comment variables so that enumerators can record additional information where necessary about why certain data is missing.

When designing data collection tools it is important to enable the enumerator to distinguish between informative non-responses and missing data, you will often have to train the enumerators to understand the distinction with respect to your specific activity.

There are many different scenarios where missing data could occur, and so to capture these reasons it is necessary to provide different codes to explain why data is missing. In many circumstances the code itself will not provide complete information about the reasons for the missing information, hence the need for additional comments variables. Suggestions for codes and examples of missing data types are outlined in Table 1.

Table 1 - Types of missing data and documentation.

| Missing Data Type | Type of Data | Example | Suggested code | Additional |
|---|---|---|---|---|
| 'Unknown' | Any | Responder does not know the answer | -99 | |
| 'Not Applicable' | Any | Field is not relevant for circumstance given previous responses | -98 | |
| 'Non Response' | Any | Responder refused to answer a question | -97 | |
| 'Non Agreement' | Any | If more than one respondent was present & they could not reach consensus on the correct response | -96 | |
| 'Not Recorded'/ 'Human Error' | Any | Field was left blank mistaken or measurement was not taken | -95 | |
| 'Technical Error' | Any | | -94 | |

| | | | | |
|---|---|---|---|---|
| 'Other' | Only Categorical | Response provided is not on the predefined list of categories. | Specify the details under 'Other', or add a code to the code list (however care needs to be taken as other enumerators should also be simultaneously adding extra codes to the same list to represent their categories). | A special, preventable case of missing data. It can be pre- empted by including a code for 'Other' in the code list, and if required a free text field for 'Specify…' |

#### 5.2.3 Multiple response data

Multiple response data usually occurs in a list format, for example, responses to questions such as: list the main types of crops grown on your land, what livestock do you keep, or what fertilizer/s have you applied to this plot. Questions such as these elicit several responses. (Alternatively, asking a series of yes/no questions -- such as do you grow crop x, crop y, etc. – would lead to single responses for each question.)

Multiple response data can be managed in two possible ways, and so thought needs to be given to this when designing the data entry software (or setting up the worksheets in software such as Microsoft Excel in which the data is to be entered):

- ✓ The number of variables (columns) that are dedicated to a multiple response question should be the maximum number of responses given by any individual. The first response given by a respondent is entered into the first variable, the second response into the second variable and so on. If a respondent only specifies 2 responses, then only the first 2 variables will have data and the other variables should be set to blank/missing.
- ✓ Another option for storing the responses to a multiple response question is to create a variable for each unique response given, and indicate whether each crop was specified by the respondent using Yes/No or coded 1/2 responses.

#### 5.2.4 Other formats of data

Data is not just items which can easily be input into spreadsheets. Data encompasses study documentation, in the form of reports, protocols, photographs, automated computer output, videos, maps, technical drawings, presentations and so on.

- ✓ Documents. Please refer to the guides on Annex 6, Data & Document Storage for details of how to store documents such as protocols, data management plans, and reports.
- ✓ Images. Images, including photographs, maps, graphs and technical drawings, should be retained in their original raw format on file. When placed into the activity archive they should be saved in an uncompressed, non-proprietary format. Depending on the size of the image the best options currently are either png (preferred) or jpeg.
- ✓ Audio/Video. All audio and video files should be retained in their original format on file. When placed into the archive they should be saved in an uncompressed, non-proprietary

format. However uncompressed video files are often extremely large in size, so for instances where many large files need to be archived compression may be necessary.

- ✓ Raw data. Raw data encompasses everything that was obtained in the data collection process, whether it is data directly gathered by the study or data that has been collected prior to the study that becomes part of the study datasets. Data can come in a wide variety of formats; some immediately useful and ready for analysis whilst others will require steps to input, format, validate and derive before any analysis can be done. The result of this process is the primary data.

## Annex 6. Data Storage and Archiving

Storing your research data is important for several reasons. First of all, according to the CIP's Open Data and Data Management Policy, researchers are obliged to store their raw research data for validation purposes and give open access to their research data.

### 6.1 Data Storage and Archiving

#### 6.1.1 Where to store

Data storage need to be well-organized throughout the project. Creating a data archive for the project is relatively straight-forward in any of the following scenarios ensuring proper backup, performance and functionality is available according to your needs:

- ✓ Work in progress data in your personal computer must be stored My Documents
- ✓ Data that needs to be shared with local HQ users must be stored in a local network shared drives
- ✓ Personal work in progress data that need to be shared must be stored in OneDrive for Business (you have up to 1 TB)
- ✓ Data that needs to be produced collaboratively must be stored in a SharePoint Team Site Library

Other personal computer folders, external storage devices such as USBs, CDs and DVDs, email, and third party online storage (including Google Drive, Dropbox and similar) are strongly discouraged for research data storage, as none of these are included in the corporate backup.

Use of corporate managed data storage allow our compliance with:

- ✓ Obligations: funder or journal requirements to make your research data openly available for re-use or validation purposes
- ✓ Value: potential value of the research data, regarding quality, originality, size, scale, collection costs, innovation
- ✓ Uniqueness: data exists of unique, non-repeatable observations
- ✓ Importance for general historic research (heritage)

#### 6.1.2 What to store

- ✓ Raw data file: the raw data file contains the originally collected, unprocessed data.
- ✓ Derived dataset: the derived dataset is the dataset underlying certain results or publications. You can derive different datasets from your raw data for different purposes.
- ✓ Syntaxes: a syntax file contains the code, algorithms or commands used to create your derived dataset from your original, raw dataset. It also contains (stepwise) information about the transformations and analyses performed on the raw dataset.
- ✓ Metadata file: a metadata file is a separate file attached to your dataset, which contains information about your dataset for future use (by yourself or others). For example, a

metadata file should contain information on the following subjects: creator, access conditions, context, collection methods, time references, structure and organization of data files, variable names, labels and descriptions of variables and values, codes for missing values, file formats, and hard- and software used to process and analyze the data. (Annex 4, Documentation & Metadata).

- ✓ As common sense dictates, storing and sharing (sensitive) data should be handled with care (Annex 7, Privacy & Confidentiality). The level of precaution that should be taken depends on the sensitivity of the data.
- ✓ Datasets underlying one or more (scientific) publications should be archived directly at CIP's Dataverse to ensure permanent preservation. In this case, you should at least store the dataset(s) used for your final analyses and results, including syntaxes that can be used to replicate your results. (Annex 8, Data Repositories)

#### 6.1.3 Preferred file formats

To ensure long-term preservation that is independent of certain specific software, you are encouraged to save your files in commonly used and easily re-usable file formats with open documentation. Please find a list of different preferred and acceptable file formats for different types of data:

| Format | File Extensions |
|---|---|
| Comma-separated values | .csv |
| Open Office formats | .odt, .ods, .odp |
| Plain text (US-ASCII, UTF-8) | .txt |
| XML | .xml |
| Shapefiles and raster files for GIS data | .shp, .tifw, .asc |
| Multimedia and pictures | .jpg |

#### 6.1.4 Reproducibility

In general, any scientific work should be reproducible. This applies to the social sciences as much as it does to the natural sciences. In practice, this means that the whole process of how you handle data should be documented. Gathering, cleaning, coding, transforming and scaling as well as analyses performed should all be documented. It is good practice to perform the above tasks using syntax, and to store the syntax along with the data.

### 6.2 Data Organization

Organizing data and documents is similar to organizing files and folders on a PC. However, for a team of researchers, it is necessary to establish a structure and naming convention for the folders and files. CIP is implementing the ISA-TAB format for files and folders organization. ISA-TAB is an abbreviation of "Investigation-Study-Assay TAB delimited format". The ISA-Tab format is a general purpose framework with which to collect and communicate complex metadata (i.e. sample characteristics, technologies used, type of measurements made) from experiments employing a combination of technologies and references to the material data files. ISA-TAB has a purpose to

capture and communicate the complex metadata required to interpret experiments employing the associated data files.

### 6.2.1. Folder Structure

The three key entities around which the general-purpose ISA-Tab format for structuring and communicating experimental metadata is built are the investigation, the study and the assay. The hierarchical structure of this format enables the representation of studies employing one or more technologies. Four types of folders are used to capture the data and metadata, and together describe a complete research project (see Figure 1).

a) The Investigation folder

b) The Study folder

c) The Assay folder

d) The Data folder

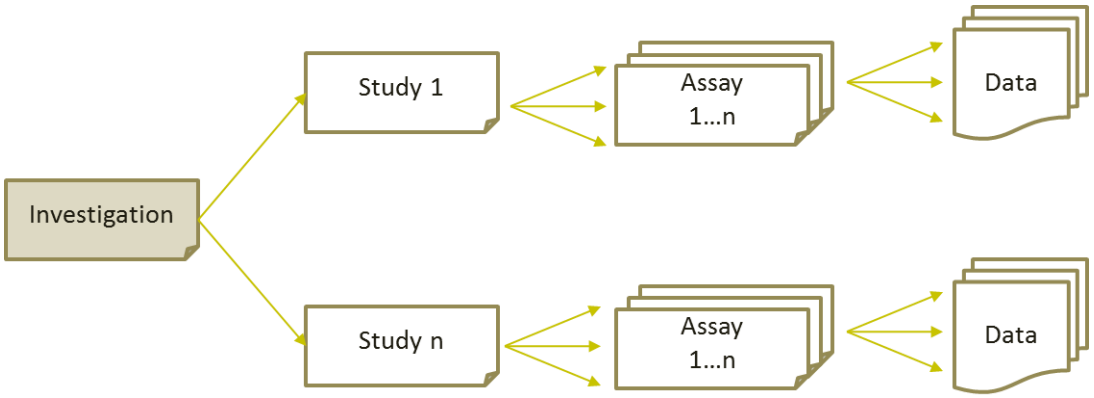


**Figure 1.** Folder structure of the research project.

The investigation folder contains all the information necessary to understand the design and overall goals of the experiment. Experimental steps (or sequences of events) are described in the study and assay folders. Each investigation folder may content one or more study subfolders; each study folder may content one or more assay subfolders and for each assay folder may content one or more data files.

**a. Investigation folder**

The investigation folder is intended to meet three needs: (i) to define key entities, such as factors or protocols, that may be referenced in the study or assay files; (ii) to track the provenance of the terminology from controlled vocabularies or ontologies and (iii) to relate assay files to study files. The minimum document requirements in this folder are:

- ✓ Research protocol (Annex 1, Research Protocols)
- ✓ Data Management Plan (Annex 2. Data Management Plan)
- ✓ Investigation Metadata file (Annex 4. Documentation & Metadata)

**b. Study folder**

The study folder contains contextualizing information for one or more assays. The minimum document requirements in this folder are:

- ✓ Manual, procedures or guidelines for the study if applicable for all assay subfolders (Annex 1, Research Protocols)
- ✓ Study Metadata file (Annex 4, Documentation & Metadata)

**c. Assay folder**

The assay folder represents a portion of the research project used to generate a dataset. Each assay folder must contain data of the same type defined by type of measurement or technology employed. The minimum document requirements in this folder are:

- ✓ Manual, procedures or guidelines for the study (Annex 1, Research Protocols)
- ✓ Assay metadata file (Annex 4, Documentation & Metadata)
- ✓ Data dictionary (Annex 4, Documentation & Metadata)

**d. Data folder**

The data folder must contain 2 types of folders. One folder must contain the raw data and the other folder must contain the final data. In exceptional cases, if the raw data is too heavy, the dataset can be stored in an institutional repositories.

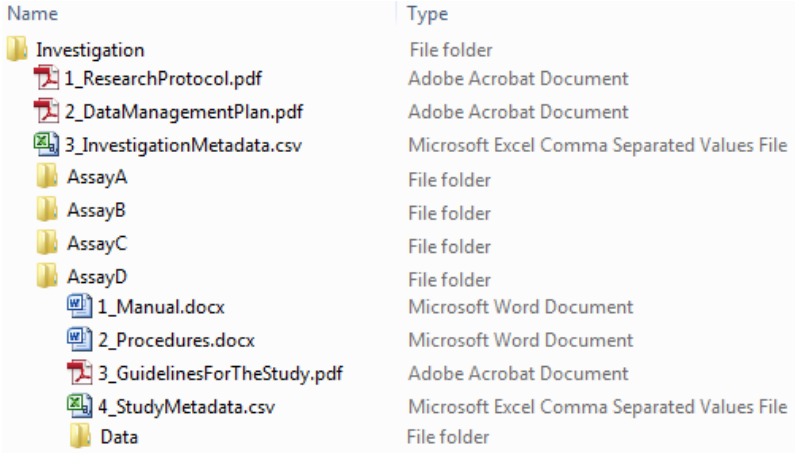


**Figure 2.** Example of how organize the research files

**6.2.2. Naming files and folders**

File naming conventions are standard methods of naming files/folders. Some considerations for file names:

- ✓ Keep file names short but meaningful (give a basic idea of its content).
- ✓ Use only alphanumeric characters.
- ✓ Avoid unnecessary repetition and redundant words.
- ✓ Use capital letters to delimit words or use underscores to separate words. Do not use spaces.

In the example file “PTYL201502_CIPHQ”, it is possible to recognize the type of trial (PTYL, potato yield), the start date (February, 2015) and the location (CIPHQ, CIP Headquarters) of the experiment.

### 6.3. Data and Documents Storage Facilities

Data and documents storage facilities need the following considerations:

- ✓ Define who is responsible for managing the Data storage at the Investigation Level.
- ✓ Follow the ISA-Tab conventions for organizing and naming files and folders.
- ✓ Give read/write permission access for the research team members.
- ✓ Train the research team members to understand how to organize data and documents.

### 6.4. Data archiving

In Research Data Management, data archiving (preservation) is the process of maintaining research data from projects so that it can still be found, understood and used in the future. Data must be kept securely even once the research has ended.

For default, CIP will retain the data for more than 10 years after finishing the project. Special considerations need to be taken into account for long-time archiving. For example, germplasm data will be retained permanently. Not all data need to be preserved for the long-term.

Before final document and data is archived need to consider the following:

- ✓ Need to include only final version.
- ✓ Exclude the unused and non-curated data.
- ✓ Compress large data; compressed files take up less disk space and download faster (suggested compression method of 7z format).
- ✓ Anonymize datasets.

The minimum document requirements for archiving include:

- ✓ The research protocol, so others can clearly see the focus of your research (Annex 1. Research Protocols).
- ✓ The Data Management Plan to define what you will do with your data during and after a research project (Annex 2. Data Management Plan).
- ✓ Metadata document for investigation, study and assay (Annex 4. Documentation & Metadata).
- ✓ Data dictionary (Annex 4. Documentation & Metadata).
- ✓ Consent agreements if needed (Annex 7. Privacy & Confidentiality).

Research data and research records must be retained for as long as required by contractual arrangements with research partners. Decisions about data retention and disposal should be documented in the data management plan, and archive with the data.

## Annex 7. Privacy and Confidentiality

Researchers are responsible for the ethical treatment of data. Research data which includes confidential or private information must be managed in accordance with any contractual or funding agreements. The researcher must seek ethical clearance at the outset of the research project. Where applicable, researchers should document:

- ✓ the nature of any private, sensitive or confidential information that may be collected
- ✓ non-disclosure agreements and any restrictions on use of the data
- ✓ consequences/penalties for breaches of confidentiality
- ✓ steps to be taken to safeguard privacy and confidentiality.

### 7.1 Privacy

Research data, particularly in social and health-related disciplines, may contain personal information about identified individuals. 'Personal' information is information that can be used to identify an individual. 'Sensitive' information includes health information, geo location information or personal genetic information and is accorded higher levels of protection than other personal information.

In research involving humans, researchers should consult the national authorities and establish conditions of use for data gathered. Researchers must comply with the national requirements to provide safeguards for the handling of an individual's personal information in the public sector environment.

Researchers must obtain consent of participants to collect sensitive information. The process involves explaining the purposes, methods, risks and outcomes of the research including the extent to which the information will be disclosed to others. Where data has been 'de-identified' (for example by replacing names with numerical IDs) it can generally be disclosed to others provided re-use falls within the nature of the purposes covered by the consent.

### 7.2 Confidentiality

A dataset or database may include information that is secret or confidential. The information can be protected from unauthorized access by use of technical mechanisms such as encryption, and passwords or legal mechanisms such as a confidentiality agreement. The actual or threatened disclosure of confidential information, which may include data that has not been made public, may result in legal liabilities.

A confidentiality agreement (also known as non-disclosure agreements) should be used where researchers wish to share confidential data on the understanding that it will not be further disclosed or used for purposes other than those covered by an agreement.

If you need to draft a confidentiality agreement for use with members of the research team, collaborators or for recipients of confidential data then contact the Research Informatics Unit's Services. You will need to supply the relevant information about data owners and the data that is to be kept confidential.

Typically, the agreement will:

- ✓ identify the owners of rights in relation to the confidential data
- ✓ describe the data that is to be kept confidential
- ✓ oblige the person to whom the data is disclosed to maintain the confidentiality
- ✓ describe the extent of any permitted use
- ✓ define the consequences of any failure to comply with the confidentiality obligations

If the data include information contributed by or about indigenous peoples then indigenous knowledge systems and processes must be respected. This includes respecting Indigenous peoples' right to maintain the secrecy of Indigenous knowledge and practices.

### 7.3 Consent agreement

Prior informed consent (PIC) is meant to guarantee the voluntary participation in research and is probably the most important procedure to address privacy issues in research. Prior informed consent consists of three components: adequate information, voluntariness and competence. This implies that, prior to consenting to participation, participants should be clearly informed of the research goals, possible adverse events, possibilities to refuse participation or withdraw from the research, at any time, and without consequences. Research participants must also be competent to understand the information and should be fully aware of the consequences of their consent. Although informed consent is often seen in the context of clinical research, this principle is important for all types of research, including the social sciences.

PIC is required in when the research involves the participation of human beings, when the research uses human genetic material or biological samples and when the research involves personal data collection. In some cases, the 'traditional' informed consent procedure might not be sufficient to ensure that the rights and interests of the research subjects are fully respected. It is very important to take into account people's autonomy and vulnerability when deciding on how to organize the consent procedure. Some categories of people require special attention:

- ✓ Children
- ✓ Vulnerable adults (elderly, prisoners, mentally deficient persons, severely injured patients …)
- ✓ People with certain cultural or traditional background: In some communities, the notion of individuality is lacking, written agreements do not exist, or women are not permitted to act autonomously. Strategies must be developed to address these issues with respect for the specificities of the situation.

The way participants are informed is a critical part of the informed consent process. When participants are informed, it is crucial that they fully realize the impact of the research for themselves and for society. Numerous anthropological studies have pointed out that participants rarely recall what they agreed upon when signing an informed consent form. A more interactive approach can address this issue. Good examples are making a presentation of the research

project or conducting interviews with the participants to ensure that they understand all the issues at stake.

PIC is an ethical requirement for most research and must be considered and implemented throughout the research lifecycle from planning to publication. Gaining consent must make provision for sharing data. Researchers must inform participants about how research data will be stored, preserved and used in the long-term and how confidentiality will be maintained. It is customary to provide an information sheet to the participants detailing the project and what their involvement will be if they agree to participate. This information sheet must cover the following topics:

- ✓ The purpose of the research
- ✓ What is involved in participating
- ✓ Any benefits and/or risks
- ✓ How the data will be used
- ✓ How the data will be stored and used in the future
- ✓ Procedures for maintaining confidentiality
- ✓ Details of the research including the funding source, who is sponsoring the project, contact details for researchers.

### 7.4 Anonymity

Before archiving data you must ensure that the dataset is anonymous – i.e. an individual cannot be recognized from their data. Obviously this would include removing names and addresses of individuals, but there are other things to take into account.

Suggestions for Anonymizing of text are:

- ✓ Don't collect personal data unless this is necessary – e.g. don't ask for full names if they can't be used in the data;
- ✓ Use pseudonyms or replacements that are consistent across the project – e.g. use the same pseudonyms in publications or follow-up research;
- ✓ Use find and replace techniques carefully so that unintended changes are not made and misspelt words are not missed;
- ✓ Identify replacements in text clearly for example using [brackets];
- ✓ Keep original versions of data for use within the research team but don't make them public;
- ✓ Create a log of all replacements, aggregations or removals; store the log separately from the anonymized data file.

In the following table, data anonymized is shown:

| Interview number | Original data | Changed to |
|---|---|---|

| | | |
|---|---|---|
| 1 | Age 54 | Age range 50 – 55 |
| 1 | 20 June | June |
| 1 | Cathy (real name) | Jane (pseudonym) |
| 2 | Station Hill Primary School | A primary school |
| 2 | Rachel | A woman |

Anonymizing audio-visual data is more difficult as obscuring faces or altering voices can reduce the usefulness of the data. If confidentiality of audio-visual data is an issue it is better to obtain the participant's consent to use and share the data unaltered.

## Annex 8. Data Repositories

### 8.1 General

Traditional approaches to storing and sharing have been either inadequate or unattractive to researchers, resulting into only a few scientists sharing their research data. Most professional archives, although often considered the most reliable solution, do not usually facilitate control and ownership of the data by the author. Once the author submits the data, the archive becomes fully responsible for the data management, cataloging and future updates. While this can be advantageous for some researchers, many prefer to maintain control of their data and to receive increased recognition. Consequently, a researcher will often choose either to offer his/her data only through his/her own website or, more commonly, to simply not share his/her data at all. Neither choice provides adequate provisions for future preservation or a permanent identifier and access mechanism. Journals and grant funding agencies are starting to require researchers to provide free access to the data underpinning their research, along with descriptive metadata, and that data is deposited and shared via suitable repositories. It is therefore becoming more important than ever to have a solution that satisfies these requirements while proving beneficial to data owners.

### 8.2 Data Repositories at CIP

Data repositories serve as platform for researchers to search, download, and analyze datasets in industry-accepted ways.

The list of CIP data repositories are:

- ✓ CIP Dataverse – is a commonly-used open source data repository platform that facilitates the ability to publish, share, reference, extract and analyzes research data. It helps to make research data openly accessible. Dataverse has been adopted at the majority of the CGIAR centers.
- ✓ BioMart database – http://cipotato.org/biomart (used to store genomic, genetic, phenotype, pedigree, geographic, environmental and experimental trials data). Biomart is already in use at CIP and will continue to be used as a complementary tool.
- ✓ RTBMaps – http://www.rtb.cgiar.org/RTBMaps (provides references, published articles and information on RTB crop layers, biotic constraints, abiotic constraints, biodiversity, socioeconomic and management of geographic information of RTB crops worldwide),
- ✓ Bioinformatics Resource Portal – http://hpc.cip.cgiar.org (CIP-HPC provides access to scientific databases and software tools, for example: genomics, population genetics, transcriptomics, and bioinformatics software),
- ✓ GitHub - https://github.com/orgs/InternationalPotatoCenter (GitHub is a web-based Git repository hosting service, from this repository is possible to download the source code packages and installers the software and algorithms development by CIP, distributed with MIT License), and

- ✓ Open access community repositories (it is an external scientific community data portals used by researchers, as in the case of NCBI-Genebank)

CIP also contributes data to external repositories, some for CRPs, such as: CCAFS Global Repository of Evaluation Trials (Agtrials), CCAFS Dataverse; and also coordinates (with partners) its own data repositories such as the Collecting Mission Fieldbook and passport repositories. Information on the material exchange through the Standard Material Transfer Agreement (SMTA) is provided to the Treaty which stores it in a non-open repository (due to confidentiality rules) for producing statistics.

### 8.3 Dataverse

CIP has implemented a locally-hosted Dataverse installation in order to customize metadata fields to be compliant with CG Core and to allow for more customization than is possible within the shared Dataverse environment (https://data.cipotato.org/). One of the reasons Dataverse has been chosen is that it can easily be linked to CGSpace and it is envisaged that users will appreciate access to datasets that form the basis of our research publications and therefore we will focus on making those links explicit where appropriate using the functionality of the two platforms.

Dataverse is a commonly-used open source data repository platform that facilitates the ability to publish, share, reference, extract and analyzes research data (King, 2007). It helps to make research data openly accessible. Each Dataverse contains studies or collections of studies, and each study contains metadata that describes the data plus the actual data (or pointers to the data) and complementary files. Additionally, each Dataverse can support harvesting of metadata to then serve as a federated search that links to other repositories hosting the actual data files.

Table 1 shows the permissions and roles in Dataverse at CIP. If someone adding a dataset to CIP Dataverse should be allowed to publish it (Admin role) or if the dataset will be submitted to the administrator of this Dataverse to be reviewed then published (Contributor role). These access settings can be changed at any time. RIU can assign those roles.

**Table 1. Permissions and roles in Dataverse at CIP.**

| Features | Admin | Curator | Contributor | Member | File Downloader |
|---|---|---|---|---|---|
| View Unpublished Dataverse | x | x | x | x | |
| View Unpublished Dataset | x | x | x | x | |
| Download File | x | x | x | x | x |
| Add Dataverse | x | | | | |
| Add Dataset | x | x | x | | |
| Edit Dataverse | x | | | | |
| Edit Dataset | x | x | x | | |
| Manage Dataverse Permissions | x | x | | | |
| Manage Dataset Permissions | x | x | x | | |
| Publish Dataverse | x | | | | |
| Publish Dataset | x | x | | | |
| Delete Dataset Draft | x | x | | | |
| Delete Dataverse | x | | | | |

Admin – RIU has all permissions for dataverses, datasets, and files.

Curator – RIU will act as a Curator for datasets, who can edit License + Terms, and then submit them for review.

Contributor - For datasets, a focal point who can edit License + Terms, edit Permissions and then submit them for review.

Member – All CIP staff and project partners (with a validate username and password) can view both unpublished dataverses and datasets, and can download file datasets. To be a member, researchers need to create an Account in Dataverse.

File Downloader – Any member of the public who can download file datasets.

### 8.3.1 Dataverse integration with CGSpace

CGSpace and Dataverse platforms were chosen as they were both recommended by the CGIAR Office as reliable tools for sharing, and promoting our research outputs, as well as having the important factors of interoperability with other systems and being OA compliant. Both CGSpace and Dataverse are used by several other CGIAR centers. Additionally, both can easily be linked to each other (See Figure 1) and it is envisaged that users will appreciate access to datasets that form the basis of our research publications and therefore we will focus on making those links explicit where appropriate using the functionality of the two platforms.

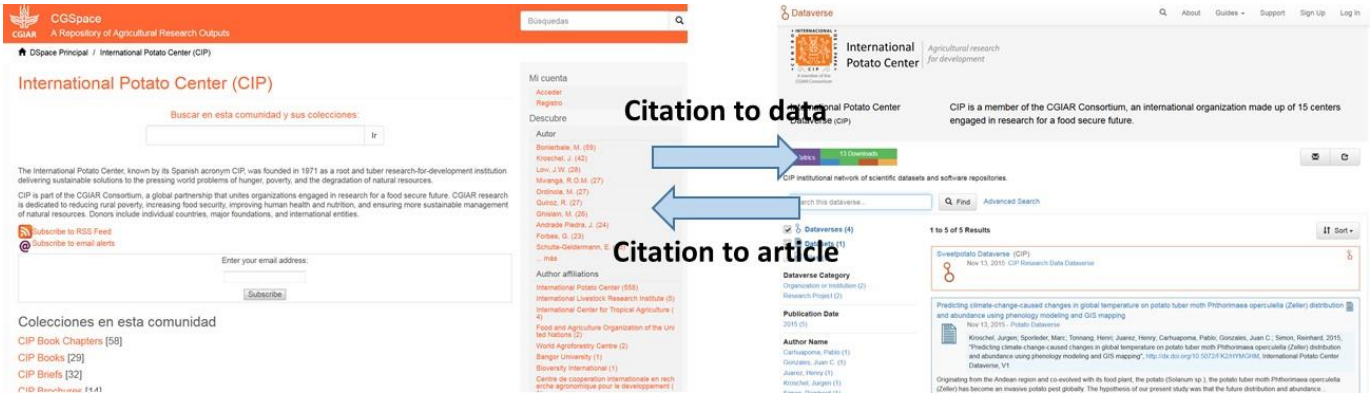


**Figure 1. CGSpace integration with the Dataverse to implement data sharing and archiving, and enhance published articles with links to data.**

### 8.3.2 Advantages to this integration

- ✓ Enabling journal editors/reviewers the ability to seamlessly manage the submission, review, and publication of data associated with published articles
- ✓ Streamlining the authors' article and corresponding data deposit process
- ✓ Permanently linking the published article with its archived data and enhancing their visibility

- ✓ Allowing authors to deposit varied data types in robust reusable preservation-friendly formats
- ✓ Ensuring that data files are discoverable, indexed, and exposed to both web and bibliographic search engines
- ✓ Enabling research data replication and reuse essential to scholarly work
- ✓ Increasing the transparency and accountability of research
- ✓ And permitting embargoes to delay release of data, in accordance with journal policy.

#### 8.3.3 How does it work?

To make data sharing and archiving as easy as possible for authors, data files are deposited in conjunction with the journal's article submission process, resulting in a permanent 2-way linking between an article and its corresponding data (See Figure 2):

- ✓ The Author submits their article to CIP library (CGSpace) and research data to RIU (Dataverse). The article and data do not have to be submitted at the same time. Authors can also submit data at a later time, or just provide a persistent link with a data citation pointing to the repository that their data is currently in.
- ✓ If the article and corresponding research data are approved for publication the Authors' research data and its corresponding metadata is automatically deposited into the Dataverse Network. No redundant information need to be entered. A permanent identifier (DOI) will be automatically included that allows the data to be cited and tracked. There will be a data citation included in the journal article page in CGSpace (and ideally within the Reference section of the article) enabling readers of the article to quickly access the data.
- ✓ The Dataverse stores the dataset metadata, files (including raw data, documentation, code, etc). There will also be a permanent publication citation link within Dataverse for researchers to access the article in CGSpace that corresponds to this research data.

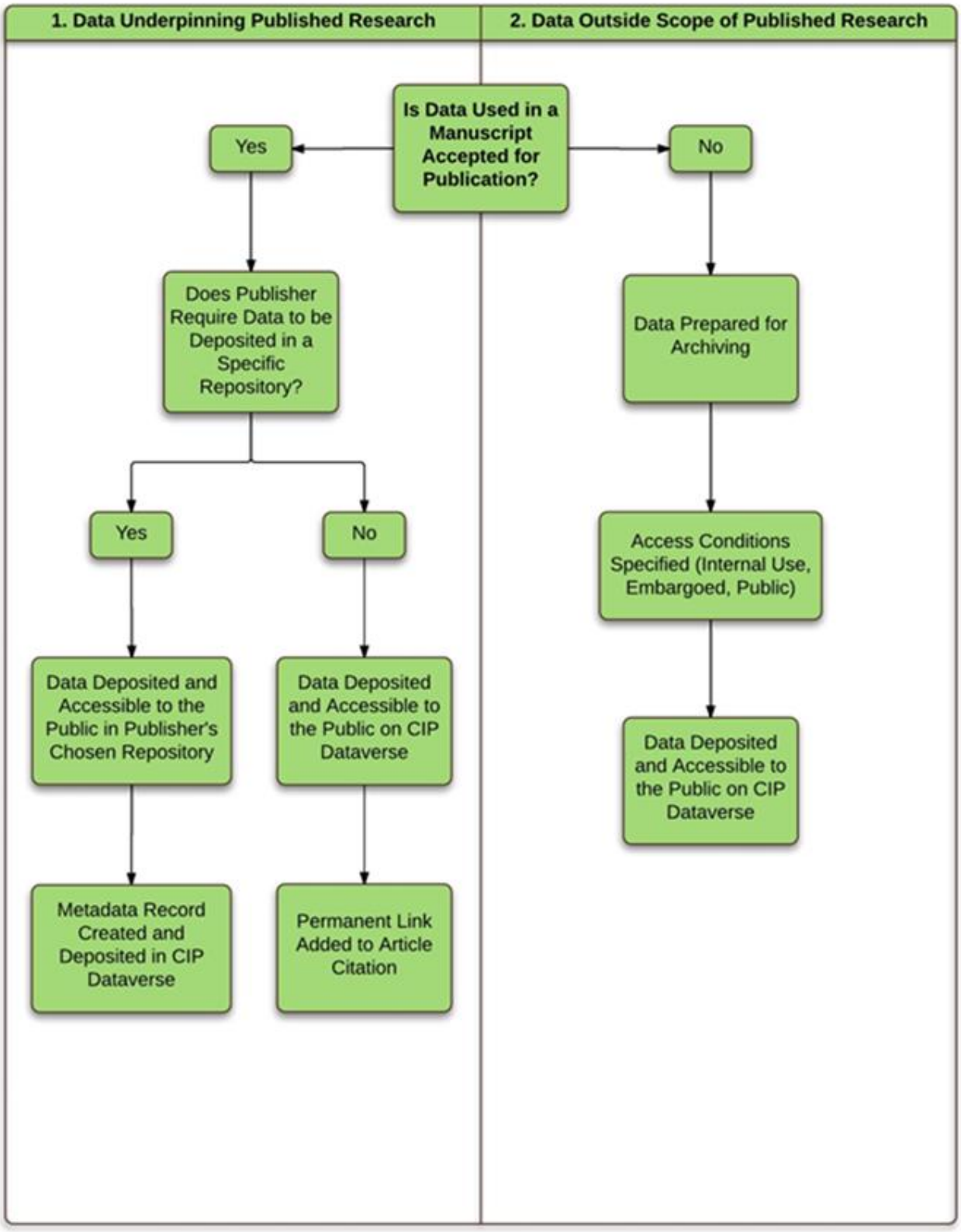


**Figure 2. Data sharing and archiving workflow resulting in a permanent 2-way linking between an article and its corresponding data.**

#### 8.3.4 Data Citation

Dataverse standardizes the citation of datasets to publish data and get credit as well as recognition for the work. When you create a dataset in Dataverse, the citation is generated and presented automatically. As an open source framework and research data repository, Dataverse is committed to help researchers, journals, and organizations make scientific data accessible, reusable, and open (when possible), which includes implementing community accepted standards for data publication.

The citation standard offers proper recognition to authors as well as permanent identification through the use of global, persistent identifiers (persistent identifier: DOI or Handle) in place of URLs, which can change frequently. CGSpace uses the data citation generated in Dataverse and add it as a reference within your publication to indicate where researchers can access your data.

The persistent URL will link to a specific dataset in CIP's Dataverse, and the dataset will contain everything needed for replication. See the following citation as example:

- ✓ Kroschel, Jurgen; Sporleder, Marc; Tonnang, Henri; Juarez, Henry; Carhuapoma, Pablo; Gonzales, Juan C.; Simon, Reinhard, 2015, "Predicting climate-change-caused changes in global temperature on potato tuber moth *Phthorimaea operculella* (Zeller) distribution and abundance using phenology modeling and GIS mapping", http://dx.doi.org/10.5072/FK2/HYMGHM, International Potato Center Dataverse, V1

In order to ensure proper data citation:

a) Add all relevant descriptive metadata in your Dataset in Dataverse.

    - Publication citation: Including a permanent link to the original publication(s) (e.g., journal article, dissertation, etc) based on the data.
    - Data Citation Details:
        - Title of the Dataset:
        - Author(s)
        - Publication date: automatically generated in Dataverse when you publish your Dataset.
        - Persistent Identifier (DOI): automatically generated in Dataverse when you create a Dataset.
    - Description and Scope
        - Description/abstract taken directly from your publication.
        - Keywords
    - Data Collection / Methodology: Add more descriptive metadata to explain how the data was collected and analyzed.

- License + Terms of Access: Copyright of all Materials created by CIP Personnel during the course of their employment or service is vested in CIP. Unless otherwise specified, CIP will adopt the 'Creative Commons – Attribution – License' (CC BY) for its copyrighted materials as well as good scientific practices require that proper credit is given via citation.

b) Add all relevant files (research data, documentation, code and analysis files) in a Dataset in Dataverse.

- List of code, scripts, documents and data files that are needed in order to make replication possible.
- Create a dataset:
  - Deposit preferred or commonly used file formats in your discipline to ensure that others will be able to more easily replicate your research. Please remember to remove information from your datasets that must remain confidential (ex. names of survey respondents). For the Social Sciences: Original SPSS, STATA, R files, csv, xslx, etc with variable names and description.
- Sets of computer program recodes (if needed).
- Program commands, code or script for analysis (if needed).
- Extracts of existing publicly available data (or very clear directions for how to obtain exactly the same ones you used).
- Documentation files (full set of supporting documentation)
  - "readme" file (explanatory document on how to use the files to replicate the study)
  - Text/pdf file of the article (if no subscription required).
  - Include a list of links to software or newly generated software used to replicate the data
  - Codebook
  - Data collection instruments
  - Summary statistics
  - Project summaries
  - Bibliographies of publications pertaining to the data

### 8.4 References

King, G. 2007. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing, Sociological Methods & Research, 36(2): 173-199.

## Annex 9. Data Management Checklist

The checklist enumerates items that are to be accomplished as part of data management throughout a project. This checklist can help you identify what to put in place for good data practices, and which actions to take to optimize data sharing. This is not an exhaustive list. The checklist template is flexible to incorporate additions or deletions for a specific project. For each item, place a check depending on whether the item has been accomplished or not.

| | |
|---|---|
| **Project Title:** | |
| **Project Lead Center:** | |
| **Project Investigator:** | |
| **Individual responsible for Data Management:** | |
| **Donor** | |
| **Agreement Id or Cost Center** | |

| **Planning (Annex 1, Annex 2, Annex 3)** | |
|---|---|
| | Is there a research protocol, budget & plan, and data management plan in place? |
| | Does the plan include the name(s) of those who are the responsible for execution of the data management plan? |

| **Documentation & Metadata (Annex 4)** | |
|---|---|
| | Are you using CIP Core metadata schema for describing the data? |
| | Are you using a data dictionary for recording variables? |

| **Data Quality (Annex 5)** | |
|---|---|
| | Are you using CIP standardized tools for collecting the data? |
| | Are you using international standards file formats to collect the data? |
| | Are there quality assurance processes for your data? |

| **Data Storage and Archiving (Annex 6)** | |
|---|---|
| | Are you using the CIP infrastructure for storage and backup? |
| | Are you using the ISA-Tab folders and files organization? |
| | Are you using the naming conventions for the project files? |
| | Are you including annotations and metadata along with the data to enable future interpretation and understanding? |

| **Privacy & Confidentiality (Annex 7)** | |
|---|---|
| | Does the data contain sensitive, confidential or personal information? |
| | If so, are there privacy considerations surrounding the ability to share/publish the research data? |
| | Have you taken the steps for protected privacy, confidentiality and anonymising data? |
| | Do you have a Prior Informed Consent agreement (PIC) in place to guarantee the voluntary participation when the research involves personal data collection? |

| **Data Repositories (Annex 8)** | |
|---|---|
| | Are you using CIP's Dataverse for data sharing? |

| | |
|---|---|
| | Are you using any other repository for data sharing? |
| | Is the metadata along with the data to enable future interpretation and understanding? |
| | Is your publication documented in CGspace linking your dataset/metadata in Dataverse? |

| **Additional observations** |
|---|
| |