

# Data Management Support Pack: Document Set

The following list is the complete set of documents making up the CIAT Data Management Pack. Below each document we have summarised the source of the document and commented on changes that need to be made and how near to completion the document is. There are also tags below each document. In black we have given the spheres (from the Data Management framework we prepared earlier) as the top level and the topics as the sub-level. In red we say whether the document is a “Main” document for researchers at the project level or whether it is a reference document. In green we say whether the document is relevant during the design phase (decisions while designing), while managing the research processes or for the delivery of the research products. Finally, in purple we say who the document is relevant for: PI, Researcher or Technician (could be more than one).

## Policy Documents

### Reference:

1. CGIAR Open Access & Data Management Policy (2013-10)
  - *Legal*
    - *Access Rights*
    - *Licensing of data and research*
    - *Open access/restrictions*
    - *Ownership of data*
    - *Partnership agreements*
  - *Strategic Planning*
    - *Data Governance*
    - *Data Ownership*
    - *Engaging and interacting with project partners (e.g. institutions, funding agencies, data providers)*
  - *Reference*
  - *When*
    - *Decisions while designing*
  - *PI, Researcher*

Please find below the document

# CGIAR Open Access and Data Management Policy (the “Policy”)

## 1. Preamble

CGIAR regards the results of its research and development activities as international public goods and is committed to their widespread dissemination and use to achieve the maximum impact to advantage the poor, especially smallholder farmers in developing countries. CGIAR considers Open Access (defined below) to be an important practical application of this commitment as it enhances the visibility, accessibility and impact of its research and development activities. Open Access improves the speed, efficiency and efficacy of research; it enables interdisciplinary research; assists novel computation of the research literature; and allows the global public to benefit from CGIAR research. Furthermore, CGIAR recognizes the benefits that accrue to individual researchers and to the research enterprise from wide dissemination, including greater recognition, more thorough review, consideration and critique, and a general increase in scientific, scholarly and critical knowledge. CGIAR further recognizes that, in implementing this Policy, it can more easily and collectively build the infrastructure necessary to be at the forefront of the open access and open data for agriculture movement.

This Policy stems from – and complies with – the *CGIAR Principles on the Management of Intellectual Assets (“CGIAR IA Principles”)*<sup>1</sup>, which is the umbrella document for this Policy. In particular, this Policy expands on Article 6.1 of the CGIAR IA Principles which provides that *“The Consortium and the Centers shall promptly and broadly disseminate their research results, subject to confidentiality as may be associated with [certain] permitted restrictions, or subject to limited delays to seek IP Rights [(patents, etc.)]”*.

## 2. Scope and implementation

This Policy was approved by the CGIAR Consortium on October 2, 2013 and is effective as of this date (the “**effective date**”). Implementation of and compliance with this Policy by the CGIAR Consortium, its members and their partners within the scope of the Strategy and Results Framework (“**SRF**”) and the CGIAR Research Programs (“**CRPs**”) will be phased over a transition period. The transition period runs from the effective date of the Policy for an initial period of 5 years, with comprehensive implementation by the end of 2018. This Policy should be read in conjunction with the CGIAR Open Access and Data Management Implementation Guidelines<sup>2</sup>, which may be updated from time to time to reflect current recommended practices.

## 3. Information products

This Policy sets common expectations with respect to Open Access to the following indicative types of information products (“**information products**”): peer-reviewed journal articles; reports and other papers; books and book chapters; data and databases; data collection and analysis tools (e.g. models and survey tools); video, audio and images; computer software; web services (e.g. data portals, modeling on-line platforms); and metadata associated with the information products above.

---

<sup>1</sup> The CGIAR IA Principles are available at <http://bit.ly/1PKeY7Z>

<sup>2</sup> The first draft will be published for consultation in late 2013, with adoption in early 2014.

## 4. Policy statement

### 4.1 General

4.1.1. Openness. Best efforts shall be used to make all information products Open Access, subject always to the legal rights and legitimate interests of stakeholders and third parties, including intellectual property rights, confidentiality, sensitivity (including price and politically sensitive information), farmers' rights and privacy.

Information products may not always be of value to others, for example because those outputs are draft, poor quality or incomplete. Open Access arrangements should consider the characteristics of the information product, their potential impact, the level of data processing required, and whether the information products generated are within the scope of this Policy. Some judgment therefore needs to be made over the information products that will be made Open Access.

4.1.2. Suitable Repositories. Stable, permanent, Open Access repositories shall be utilized, to enable users and other sites and search engines to access or locate information products, including application programming interfaces (APIs) or other mechanisms enabling those information products to be available from the CGIAR website and associated web-based products. Preference should be given to existing repositories to minimize the number of repositories in use (and the interoperability challenges presented by multiple incidences of repositories).

4.1.3. Interoperability. Syntactic and semantic interoperability is a key consideration in enabling and promoting international and interdisciplinary access to and use of information products. Information products must therefore be described with standardized metadata, and stored and delivered using appropriate protocols and formats to ensure that their content can be discovered, shared and incorporated across different technological platforms.

4.1.4. Data storage and preservation for future use. Information products must be stored where users can find them and where they will be preserved for future use. As time goes by, they will need to be managed, maintained and curated.

4.1.5. Copyright and Open Licenses. Suitable open licenses shall be used that recognize the legal rights to information products and encourage their use and adaptation.

4.1.6. Incentives and professional expertise. Incentives and the development of professional expertise in all areas of Open Access and data management shall be devised, adopted and promoted.

4.1.7. Translation. Translations of key documents and other media into pertinent languages are encouraged. All versions should be deposited in suitable repositories and made Open Access.

4.1.8. Limited internet connectivity. To assist those with limited internet connectivity, designing easily accessible information products (e.g. websites, PDFs) or making available alternative versions that require minimal data download to see and use is encouraged.

4.1.9. Open Access and Data Management Plans. Open Access and Data Management Plans should be prepared in order to ensure implementation of this Policy. Such Plans shall, in particular, outline a strategy for maximizing opportunities to make information products Open Access.

## 4.2 Open Access for indicative types of information products

4.2.1. Peer-reviewed journal articles. Peer-reviewed versions of scholarly articles reporting research should be deposited in a suitable repository and made Open Access as soon as possible, ideally at the time of publication, and no later than 6 months from the date of publication. Authors are free to choose the journal that is most appropriate to their needs. Where an author publishes in a closed access journal, he/she shall self-archive in an Open Access repository a digital version of the final accepted manuscript (the “postprint” version).

4.2.2. Reports and other papers. Information products that are not intended for peer-review journals, such as reports, conference papers, policy briefs and working papers, shall be deposited in suitable repositories and made Open Access as soon as possible and in any event within 3 months of their completion.

4.2.3. Books and book chapters. The full digital version of books and book chapters shall be made Open Access as soon as possible after publication and in any event within 6 months either through self-archiving or other suitable publication arrangements.

4.2.4. Data and databases. Data (and any relevant data collection and analysis tools) shall, subject to any additional donor requirements, be deposited in a suitable repository and made Open Access as soon as possible and in any event within 12 months of completion of the data collection or appropriate project milestone, or within 6 months of publication of the information products underpinned by that data, whichever is sooner. Data deposited shall be prepared in a manner consistent with the aims of this Policy. Existing and future databases shall be made Open Access.

4.2.5. Video, audio and images. Complete final digital versions of video and audio outputs, and image collections must be stored appropriately and made Open Access within 3 months of their completion.

4.2.6. Computer software. Where an information product is software developed internally, the associated source code must be deposited in a free/open software archive upon completion of the software development. Access to such information products may be granted subject to appropriate licences (e.g. Copyleft).

4.2.7. Metadata. The metadata of an information product must be deposited in a suitable repository before or on publication of the information product. Where an information product is not deposited in a suitable repository, the deposited metadata must include a link to the information product.

## 5. Review

The CGIAR Consortium Office will carry out an evidence-based review of the implementation of this Policy on an annual basis. The reviews will be used to devise appropriate institutional tools and guidelines for the implementation of this Policy.

The Consortium Office (in consultation with the Centers) will review this Policy in 2015 and every two years thereafter in light of experiences gained. This Policy may be amended at any time by agreement of the Consortium, in consultation with the Centers.

## 6. Definitions

For the purposes of this Policy:

**Data** means the digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings, including data sets used to support publications and/or that have been prepared and validated but that do not support publications. This does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as laboratory specimens.<sup>3</sup>

**Database** means a collection of independent works, data or other materials, which are arranged in a systematic or methodical way and which are individually accessible by electronic or other means.<sup>4</sup>

**Open Access** means the immediate, irrevocable, unrestricted and free online access by any user worldwide to information products, and unrestricted re-use of content (which could be restricted to non-commercial use and/or granted subject to appropriate licences in line with the CGIAR IA Principles), subject to proper attribution.

---

<sup>3</sup> Adapted from the Office of Science and Technology Policy Guidelines.

<sup>4</sup> From Directive 96/9/EC of the European Parliament of 11 March 1996 on the legal protection of databases.

## 2. CGIAR Open Access & Data Management Implementation Guidelines (2014-07)

- *Strategic Planning*
  - *Data Governance*
  - *Project level Data Management and Sharing Strategy/Plan*
- *Data Management*
  - *Data Storage*
- *Managerial*
  - *Informed selection of data collection tools/methods*
- *Reference*
- *When*
  - *Decisions while designing*
- *PI*

Please find below the document

# Open Access & Data Management Implementation Guidelines 2014

## Section 1: Introduction

### 1.1: About the Guidelines

In November 2013, all 15 members of the CGIAR Consortium unanimously endorsed the Open Access and Data Management Policy (the “Policy”)<sup>1</sup>, a policy designed to make final CGIAR Consortium information products – including publications, datasets, and audiovisual materials – Open Access (“OA”).

The Open Access Guidelines (“OA Guidelines”) address a first phase of Consortium-wide implementation, one focused on **planning** and **coordinating** efforts throughout the CGIAR Consortium, to a point where all members of the CGIAR Consortium (“Centers”) have active Open Access implementation plans in place. The goal is to provide guidance and support to make this first milestone possible by the end of 2014. Furthermore, the OA Guidelines are premised on the following key principles which emerged during consultation with Research Centers in the months since the Open Access and Data Management Policy was passed:

- The primary responsibility for implementing Open Access must lie with the Centers.
- The Consortium Office’s key contribution will be the overall policy, supported by suggested, generic implementation guidelines that describe the basic principles to be followed in Center-developed implementation plans.
- Centers should develop their own specific and verifiable implementation guidelines and plans, according to the Consortium Open Access and Data Management Policy and this set of generic guidelines. Such plans should be tailored by each Center to their specific research programs, resources, and previous experience with Open Access and Data Management.

Thus, these OA Guidelines are intentionally broad in nature and are designed to offer as much flexibility as possible for Centers in planning for and preparing their own implementation plans and general approaches to supporting Open Access and Data Management. Even so, Centers’ plans – and, in particular, the resulting infrastructure – need to meet certain minimum criteria in order to work together across the Consortium.

Since the OA Guidelines are so broad, more detailed guidance and additional (optional) recommendations will be produced and offered through an evolving set of online resources. These resources, which will collectively be referred to as the “OA Support Pack,” will be a combination of new resources produced in response to requests for guidance by Research Centers; adaptable or re-usable materials developed by Research Centers; and materials produced by external organizations. The Support Pack will be available via <http://open.cgiar.org> by September 30, 2014.

Two notes regarding the scope of these Guidelines:

1. **Final information products.** The emphasis of the Open Access and Data Management Policy is on final research outputs – those information products (regardless of format) that are “stable” and unlikely to undergo further change (e.g., post-publication materials, datasets collected over the life of a project that has ended, etc).
2. **Data Management and Open Data.** These Guidelines focus on data within the context of Open Access – in other words, making final versions of research outputs (including data sets, analysis tools, survey instruments, models, summaries of data, maps and spatial products) openly and freely accessible for use and re-use by others. Data Management is addressed in as much as it affects making data OA.

---

<sup>1</sup> CGIAR Open Access and Data Management Policy is available at <http://bit.ly/1g8qioD>.

## Section 2: Roles & Responsibilities

### 2.1: Members of the CGIAR Consortium (the Centers)

CGIAR Consortium Centers have primary responsibility and accountability for day-to-day implementation of Open Access – for instance, establishing Center-specific policies and procedures; managing and curating repositories; establishing policies and procedures for scientists that encourage, support, and reward deposits of research outputs and data in appropriate repositories; ensuring that Center-specific implementation efforts are consistent with the Open Access & Data Management Policy; and reporting metrics via agreed-upon channels.

As an early part of the implementation process, Centers should each prepare an Open Access and Data Management Plan. Section 3 of these Guidelines provides an overview for developing these Open Access and Data Management Plans; further details are included in the Annex.

CGIAR Research Program Lead Centers are responsible for ensuring that program participants and partners are compliant with the Open Access & Data Management Policy and Guidelines through their Program Participant Agreements (PPAs). CRP Lead Centers must appropriately allocate sufficient resources to allow for the implementation of Open Access. CRPs may choose to unilaterally adopt the Lead Center's Open Access plan, or to develop their own, provided it adheres to the core basic requirements of the policy and these guidelines.

Furthermore, Centers are responsible for ensuring that all relevant future agreements and contracts, including confidentiality, partnership, collaboration, development, licensing, distribution, material transfer agreements, employment contracts, and grants, comply with the Policy and with the CGIAR Intellectual Assets Guidelines.<sup>2</sup>

### 2.2: Consortium

The CGIAR Consortium, on behalf of the CGIAR Consortium members (the Centers), is focused on advancing Open Access and Open Data at a system level, providing Centers and CRPs with the policy framework for Open Access and Data Management, as well as the system-wide coordination, aggregation, and sharing of tools, seed funding, resources, and advocacy needed by members of the Consortium to implement OA and data management.

The Consortium serves in a consultative capacity by facilitating meetings, bringing in external resources, and acting as a champion and advocate for Open Access and Open Data internally and externally as CGIAR strives to become a leader in open knowledge for agriculture research.

The Consortium shall:

- Comply with the CGIAR Open Access and Data Management Policy, including developing a Consortium Office Open Access and Data Management Plan for its own outputs and products that could serve as a template for other CGIAR entities;
- Facilitate the process to develop, maintain, assess, and revise the Open Access and Data Management Policy and Implementation Guidelines;
- Facilitate the process to develop, maintain, assess, and revise the CG Core Metadata Schema and relevant vocabularies adopted at the Consortium level;
- Develop, compile and aggregate data, benchmarks, milestones, and metrics at the Consortium level;

---

<sup>2</sup>See the Implementation Guidelines for the CGIAR IA Principles on the Management of Intellectual Assets (<http://bit.ly/1nkHhdC>) for more details.



- Compile and make available tools, resources, and content for the Open Access Support Pack;
- 
- Develop new tools to bring together content or harvested metadata across CGIAR repositories (and partner repositories, when/where applicable);
- 
- Develop new tools to analyze or visualize aggregated data from repositories across the CGIAR Consortium (including support for other value adding products and services derived from or leading to better use of open data and content);
- 
- Provide aggregated reporting to Open Access funders and other interested parties;
- Highlight and showcase emerging best practices from within the Consortium;
- 
- Negotiate contracts and Memoranda of Understanding related to Open Access that Centers might benefit from (e.g., publisher agreements that minimize OA fees);
- 
- Promote CGIAR OA principles in relevant initiatives and organizations related to Open Access and Open Data, such as AATP, GODAN, and others; and
- 
- Communicate OA activities with relevant Center staff, involving them in negotiations, priority setting, product testing etc.

### 2.3: Open Access Implementation Working Group

The Open Access Implementation Working Group (OAIWG) has been established to help create the enabling environment for Open Access implementation<sup>3</sup> and consists of Knowledge Managers from Centers, CRPs and CO representatives. It will help to oversee and guide the implementation of the CGIAR Open Access Policy between 2014 and 2018 and for managing appropriate communications around this Policy. See the Terms of Reference for the OAIWG for full details.

### 2.4: The Data Management Taskforce

The CGIAR Data Taskforce (DMTF) will take a lead role in coordinating Open Access implementation with a focus on data management issues.<sup>4</sup> The DMTF is made up of Data Managers or their equivalent from Member Centers and CRPs, along with CO representatives. In particular, the Data Management Taskforce will provide oversight of data standards and protocols and will be responsible for defining the appropriate standards and interoperability protocols to be implemented and applied across CGIAR Open Access repositories. See the Terms of Reference for the DMTF for full details.

### 2.5: CGIAR Partners

Information products produced by lead Centers and participating Centers (including partners) in CRPs are subject to the Policy on all new contracts established since the adoption of the Policy. A contract may have been entered into which contains restrictions on, for example, sharing the data under a research and/or development project or under a commercialization endeavor. Future agreements should be carefully negotiated to ensure that any such restrictions are limited in duration, territory and/or field of use, if applicable, and fully justifiable by reference to the CGIAR Principles on the Management of Intellectual Assets (i.e., in particular articles 6.2, 6.3 and 6.4)<sup>5</sup> and these Guidelines. For more details on this, work with your Center's IP focal point.

---

<sup>3</sup> See the Terms of Reference for the CGIAR Open Access Implementation Working Group for additional details.

<sup>4</sup> See the Terms of Reference for the CGIAR Data Management Task Force for additional details.

<sup>5</sup> Article 6.2 is on "Limited Exclusivity Agreements"; Article 6.3 is on "Incorporation of Third Party Intellectual Assets" and

### Section 3: Open Access and Data Management Plans & Implementation

According to Article 4.1.9 of the Open Access and Data Management Policy: *“Open Access and Data Management Plans should be prepared in order to ensure implementation of this Policy. Such Plans shall, in particular, outline a strategy for maximizing opportunities to make information products Open Access.”*

Each CGIAR Consortium Member Center should develop a Center-specific Open Access and Data Management Plan detailing how Open Access and data management will be implemented and supported. Plans should address all of the elements specified by the Policy. **Likewise, all aspects of implementation – particularly in terms of the technical infrastructure including repositories, metadata, and interoperability – should comply with the minimum parameters set forth in the Policy and these Guidelines.** See Table 1 for a high-level overview of these minimum parameters and the Annex for details. Numbers in parentheses refer to articles from the CGIAR Open Access and Data Management Policy.

Centers are expected to, at minimum, adopt, and ideally, enhance and exceed these requirements. Open Access Plans should be updated regularly to reflect current thinking as Centers move forward with implementation. More detailed guidance and examples will be forthcoming as part of the OA Support Pack.

**Table 1: Essential Elements to Include in Open Access Plans**

Essential Elements for Plans	Minimum Requirements
<b>Scope</b>	
Openness (4.1.1)	General statement regarding interpretation of openness, what is covered in the plan, brief synopsis of current OA situation/needs/challenges, how any repositories or exchange systems include partner contributions and provide them access, areas to address in the future.
Information Products (4.2)	Plans should address how all types of information products addressed in the Policy are collected, stored, and disseminated; differences in treatments based on type of information products and format of information products. Types of information products to be addressed include: peer-reviewed journal articles; reports and other papers; books and book chapters; data; data summaries (e.g. maps, indicators); video, audio, and images; computer software and models; and associated metadata.
<b>Technical Infrastructure</b>	
Suitable repositories (4.1.2)	Repository systems should meet current industry standards for interoperability and metadata. Plans should address which repositories are in use for each type of information product (see below), associated URL(s) for the repositories, and identify which repository platforms/systems are in use.  Where information or data is stored or published in systems that may not be ‘repositories’, plans should show how these meet Open Access principles and requirements, including standards of interoperability and use of metadata.
Interoperability (4.1.3)	Plans should address which interoperability protocols and standards are adopted in any repositories as well as any other digital or web-based aggregation or harvesting systems and services; how content is transferred or interpreted between systems (internally or externally); and any new or emerging tools, protocols, or frameworks being tested or considered for adoption.

Metadata (4.1.3)	Centers and CGIAR system units should demonstrate that they are working towards adopting the CG Core Metadata Schema. Plans should confirm that they have adopted CG Core and offer details on any additional metadata schemas and taxonomies which are in use and how they are applied.
Limited internet connectivity (4.1.8)	Plans should address how the Center and system units are providing access to content when internet access is limited.
Data storage and preservation for future use (4.1.4)	Plans should address storage and preservation issues such as storage redundancy; storage formats and preservation mechanisms; use of persistent identifiers; recommended & accepted file formats.
Copyright and open licenses (4.1.5)	Plans should address recommendations on which open licenses to use for different types of information products or if the Center adopts recommendations offered by the CO.
Translation (4.1.7)	Plans should address how open licenses allow for re-use such as translations of information products and ways in which translations are encouraged.
<b>Administration of Open Access &amp; Data Management<sup>6</sup></b>	
Strategy and implementation	Plans should address short-term and long-term strategic goals, objectives, and priorities for the Open Access and data management programs as well as implementation timelines.
Processes and workflows	Plans should detail workable processes to acquire and deposit material into repositories, manage/maintain repositories, assure quality control etc.
Staffing	Plans should address which department(s) or position(s) have day-to-day responsibility for Open Access and data management, which group(s) have oversight for Open Access and data management, and the general structure of these groups. Plans should address recommended workflows for depositing materials into repositories.
Financial administration	Plans should address financial administration, funding, and major expenditures for supporting Open Access and data management support.
<b>Assessment, Impact, and Review</b>	
Incentives and professional expertise (4.1.6)	Plans should address ways in which creators of information products are encouraged to comply with the Policy and how compliance is tracked, measured, and reviewed.
Assessment and review (5)	Plans and progress should be updated on a yearly basis. Plans should address mechanisms for internal review.
Tracking impact & uptake (5)	Plans should address which metrics are being collected and how they are interpreted in order to understand usage, impact, and uptake of materials disseminated through Open Access.

Content should be deposited in full as soon as possible after an item is complete or in its final form. Plans should address timelines for depositing information products into repositories. Plans' deposit schedules should be consistent with these guidelines or indicate shorter timelines than those presented here. See **Table 2** for details **Error! Reference source not found.**

<sup>6</sup> Strategy, staffing, and financial administration are not directly referenced in the OADM Policy, but are necessary for successful execution of Open Access and data management and so should be addressed in Centers' Plans.

**Table 2: Deposit Schedules**

<b>Types of Information Products</b>	<b>Deposit Schedule</b>
Peer-reviewed versions of journal articles	Ideally, at the time of publication Latest: 6 months from publication <sup>7</sup>
Self-published journals, books, reports etc.	Immediately
Reports and other papers	As soon as possible Latest: within 6 months of completion
Externally or commercially published books and book chapters	As soon as possible Latest: within 6 months of completion
Data and data sets	As soon as possible Latest: within 12 months of completion of data collection or appropriate project milestone, or within 6 months of publication of the information products underpinned by that data <sup>8</sup>
Video, audio, scientific images	As soon as possible Latest: within 6 months of completion
Photographs	As soon as possible Latest: within 6 months of completion or publication
Computer software/applications/code	As soon as possible Latest: within 6 months of completion
Metadata	As soon as possible Latest: before or on publication of the information product
Core/corporate governance documents appropriate for public consumption (e.g., financial reports, board agendas and minutes, annual reports, as appropriate)	As soon as possible

**Automated exceptions/extensions.** Certain types of information products (in particular data collected pursuant to hypothesis-driven research) may take longer than 12 months to clean, analyze and publish. Thus, 12 months should be seen as the aim, with 24 months as the long-stop date for making such data Open Access. During the implementation phase, the best timing of disclosure of information products will be identified.

**Exceptions.** The general principle is to make information products Open Access, but that is always “*subject to the legal rights and legitimate interest of stakeholders and third parties, including intellectual property rights, confidentiality, sensitivity (including price and politically-sensitive information), farmers’ rights and privacy.*”

Exceptions also include aspects referred to in Articles 6.2, 6.4, and 6.4 of the CGIAR IA Principles.

**Effective date.** The Policy is effective as of 02 October 2013. Only final information products produced after the effective date are covered under the Policy.

<sup>7</sup> Researchers are strongly encouraged to work with publishers to secure appropriate permissions. However, if a publisher’s contract prevents compliance, appropriate copies should be deposited into the repository immediately, but be blocked from public access until after the embargo is lifted. Furthermore, if an article is not able to be deposited into a repository due to a publisher’s restrictions, a metadata record should be added to the appropriate CGIAR repository and should include a link to the publisher’s website.

<sup>8</sup> The timeframe for dissemination of data is as soon as possible after data collection has been completed, and in any event within 12 months, although this may vary according to the type or nature of the results, among other factors.

## Section 4: Next Steps

### 4.1: Phase 1 – Planning and Coordinating

The emphasis for 2014 will be on planning and coordinating efforts across Centers. Much of this work is focused on collecting and aggregating information to discover what is already in place throughout the Consortium and bringing all Centers up to minimum compliance with the Policy.

Activities included in Phase 1 for 2014:

- Gather baseline information about current practices: inventory established repositories, adopted interoperability protocols, Centers' focal points. (March – May)
- Form two new groups designed to support Open Access implementation: Open Access Implementation Working Group (OAIWG) and Data Management Taskforce (DMTF) (March-May)
- Share baseline information about the current status of Open Access and Data Management as gathered via the Inventory (May/June)
- Through the OAIWG and DMTF, set minimum requirements for metadata for CG repositories (June/July)
- Begin sharing detailed guidance on a variety of topics through the OA Support Pack (June - September)
- Begin work on an advocacy plan and advocacy materials (June-September)

Furthermore, all Centers should be starting to work with researchers to collect and describe final information products. Ideally, these products will be immediately deposited into suitable repositories. For Centers without such repositories in place at this point, information products should still be collected and described. Workflows can be established even before repositories are in place.

In order to advance progress with Open Access as quickly as possible and have Members in compliance with the Policy, Centers and the Consortium should aim to complete the following tasks by these deadlines:

Activity	Deadline
Report baseline metrics via the CGIAR inventory of current practices (survey)	30 April 2014
Develop and submit Center-based Open Access and Data Management Plans	15 December 2014
Launch suitable Open Access repositories – or adapt existing repositories	15 December 2014
Incorporate metadata recommendations into new/existing repositories	15 December 2014

### 4.2: Future Phases

Once all Centers have developed their Open Access plans, have established workflows in place to acquire and deposit content into repositories, and have suitable repositories in place, Consortium-level efforts will transition to aligning the infrastructure across Centers, increasing the visibility and usage of information products, ensuring compliance of the Policy, and assessing the impact of the Policy.

## ANNEX: Guidelines for Open Access Plans

## Guidelines for Open Access Plans

Since CGIAR will be implementing Open Access in a distributed environment, it is important that each the approach meets the minimum criteria outlined in these guidelines, and is in line with resource levels, technical capabilities and infrastructure, and the exact nature of its information products.

Plans will be shared with the Consortium Office and the OAIWG and DMTF in order to foster the exchange of knowledge and spark new ideas for implementation. Plans should also be made publicly available in order to support knowledge exchange within the broader agricultural research community.

Plans should be reviewed and updated yearly to reflect progress, the adoption of best practices, and the incorporation of new technologies. The first round of implementation plans, will be due to the Consortium Office in December 2014. The Consortium Office plan is envisaged earlier, to provide an option for Centers to use it as a template. For the first phase of implementation, plans will be used as the reporting mechanism to help establish initial progress.

Furthermore, plans will be used to inform review of the Open Access and Data Management Policy and its implementation. According to Article 5 of the Policy: *“The CGIAR Consortium Office will carry out an evidence-based review of the implementation of this Policy on an annual basis. The reviews will be used to devise appropriate institutional tools and guidelines for the implementation of this Policy.”* Plans for the review will be developed in the coming months.<sup>9</sup>

Each plan should include a general statement regarding any Center--specific interpretation of openness, current OA situation, workflows to address content deposition into repositories or other systems, maintenance of these platforms, quality control etc., what is covered in the plan, and areas included in the policy that are not yet supported by the Center.

The following types of information products should be addressed in each plan:

- Peer-reviewed journal articles (Article 4.2.1)
- Reports and other papers (Article 4.2.2)
- Books and book chapters (Article 4.2.3)
- Data and databases (Article 4.2.4)
- Video, audio, and images (Article 4.2.5)
- Computer software/applications/code (Article 4.2.6)
- Metadata (Article 4.2.7)
- Core/corporate governance documents appropriate for public consumption

Some Centers might elect to support different types of information products in different ways. For instance, a Center might use DSpace to collect, archive, and disseminate peer-reviewed journal articles, Dataverse for socioeconomic datasets, and Github for computer software. Other Centers might use a single repository and offer the same services for all openly-accessible information products. At a minimum, plans should include an overview of how the different types of information products are supported.

*Additional guidance for specific types of information products will be addressed through the Support Pack.*

---

<sup>9</sup> As part of the Implementation Roadmap, the Open Access Implementation Working Group and the Data Management Taskforce will both be involved in helping to devise the evidence-based review of implementation and establishing M&E metrics.

## A: Technical Infrastructure

In order to comply with the CGIAR Open Access and Data Management Policy, each CGIAR Consortium Member Center and system unit must use suitable, interoperable, standards-compliant Open Access repositories in order to provide access to the body of information products and associated metadata produced by members of its community. Many of the Centers and units already have their own repositories – and some Centers have multiple repositories adopted for different purposes – leading to a complex, distributed network of repositories throughout the CGIAR community.

In terms of the technical infrastructure, plans should address:

- Repository systems (Article 4.1.2)
- Interoperability (Article 4.1.3)
- Metadata (Article 4.1.3)
- Limited internet connectivity (Article 4.1.8)

The Technical Requirements included here are the minimum necessary requirements in order to create an interoperable network of repositories and present open knowledge in a coherent, meaningful way across repositories, Research Centers, and CRPs. Furthermore, adoption of widely-used systems, protocols, and standards will make it possible to continue to enhance the infrastructure in the future and incorporate new tools designed to maximize visibility, discoverability, re-use, and uptake of CGIAR information products.

### A.1: Suitable Repositories

According to Article 4.1.2 of the CGIAR Open Access and Data Management Policy: *“Stable, permanent, Open Access repositories shall be utilized, to enable users and other sites and search engines to access or locate information products, including application programming interfaces (APIs) or other mechanisms enabling those information products to be available from the CGIAR website and associated web-based products. Preference should be given to existing repositories to minimize the number of repositories in use (and the interoperability challenges presented by multiple incidences of repositories.)”*

All of the information products covered in the Policy must be deposited into an Open Access repository which will then provide the necessary infrastructure to archive and disseminate the body of knowledge captured in CGIAR’s information products.

System design must be scalable, flexible, and facilitate extraction of data in multiple formats and for a range of uses as internal and external needs change, including potential uses not accounted for in the original design. In general, this will involve the use of standards and specifications in the system design that promote best practices for information sharing, and separation of data from the application layer to maximize data reuse opportunities and incorporation of future application or technology capabilities, including the ability to export information in RDF.

At a minimum, all repositories should be OAI-PMH compliant or expose metadata through standard APIs.

#### **Publication and General-Purpose Repositories**

Many Centers and CGIAR units are already using DSpace for publication repositories. DSpace is an open-source, standards-compliant, widely-adopted repository platform which meet the specified criteria. Centers or units not wishing to maintain their own publication repositories, or wishing to collaborate and cost-share should consider using



CGSpace<sup>10</sup>, a shared installation of DSpace that serves several CGIAR entities and CRPs, or hosting their own installation of DSpace. Other possibilities and recommendations will be provided by OAIWG if any Centers do not expect to have a workable repository in place within the next few months.

### **Data Repositories**

At the November 2013 CGIAR Data Standards Summit, 5 main types of data related to agricultural research were identified: (1) socio-economic; (2) spatial; (3) genetic; (4) genomic; and (5) germplasm. A major data type of relevance to Centers that is missing from these streams is agronomy trial data (particularly that focusing on management). Effective immediately, the Data Standards Task Force will begin working to coordinate and standardize data management practices around these areas, recognizing that in the NCBI suite of platforms, there are already good options for genetic and genomic data repositories in particular. Data from each of these areas has highly-specialized characteristics and will require different types of stewardship, and, in many cases different platforms.

Many Centers are already using DSpace (primarily publications), Dataverse (primarily socioeconomic data), or AgTrials (primarily breeding and agronomy trial data) as repositories. Genetic and genomic data from Centers is also deposited into NCBI databases. All these repositories are or will soon be open-source, standards-compliant, and widely-adopted systems which meet the specified criteria.

Individuals supporting data repositories are strongly encouraged to work through the Data Management Taskforce when encountering new data management challenges. Since research programs and subject areas touch multiple Centers, it is expected that others are encountering – or will encounter – similar challenges. Moving forward, it would be beneficial for Centers to work together to support emerging areas (e.g. the need to effectively archive and make accessible agronomy trial datasets).

*At a minimum, plans should:*

- Address which repositories/platforms are in use and for what purposes they are being used – e.g.: Dataverse for primarily socio-economic data sets, DSpace for peer-reviewed and other publications, Open Journal System (OJS) for CGIAR-published journals; and
- Include URLs to the repositories.

*Further guidance on repository systems and data management will be forthcoming as part of the Support Pack.*

### **A.2: Interoperability**

As a result of this distributed environment, repositories must be syntactically and semantically interoperable. This means that at a minimum, repositories should all be OAI-PMH compliant and comply with the CG Core metadata schema. Other considerations for using standard APIs can be considered by the Data Management Taskforce.

Interoperability is possible by adopting commonly-implemented technical protocols, standards, and vocabularies. Although the interoperability landscape continues to evolve, using widely-adopted repository systems, metadata schemas, ontologies, and vocabularies make it easier to incorporate new tools, protocols, and initiatives as they are released.

---

10 CGSpace home: <https://cgspace.cgiar.org/>. CGSpace is a collaboration of several Centers and CRPs and is hosted by the International Livestock Research Institute (ILRI).

At a minimum, plans should address:

- Which mechanisms are in place to enable cross-system transfer of content or metadata and how these mechanisms are being used;
- How content is transferred between systems (internally or externally); and
- Any new or emerging tools, protocols, or frameworks being tested or considered for adoption in the coming 1-2 years.

Examples of interoperability protocols or tools used in connection with Open Access and Data Management include:

- AGROVOC Linked Open Data API to use AGROVOC terms from within DSpace
- OAI-PMH enabled to allow harvesting of metadata by OAlster

*Further guidance on interoperability and examples of protocols will be forthcoming as part of the OA Support Pack.*

### A.3: Metadata

The CG Core Metadata Schema (CG Core)<sup>11</sup> will be a common framework for CGIAR Consortium Member Centers, CRPs, and other entities to present and share metadata in consistent ways across the network of CGIAR repositories. CG Core is based on Dublin Core, a widely-used metadata standard, with a limited number of additional elements specific to the CGIAR environment.

All CGIAR Consortium repositories should adopt the CG Core. It is not intended to replace existing metadata schemas used by Centers, which include additional domain-specific details and vocabularies. It is expected that most repositories will include a crosswalk to map existing element sets to CG Core elements.

Plans should confirm adoption of the CG Core elements within repositories. In addition, plans should address:

- Other metadata schemas in use and how they are applied, and
- Vocabularies in use and how they are applied.

*See “CG Core Metadata Schema & Guidelines” for current details and guidelines of CG Core. Further guidance on applying CG Core will be forthcoming as part of the OA Support Pack.*

### A.4: Limited Internet Connectivity

Research Centers are encouraged to design repositories and websites in ways that support low-bandwidth and mobile users, and others with limited internet connectivity, without compromising quality of the information products. At a minimum, plans should address how Centers support individuals with limited internet connectivity and steps taken to optimize for low-bandwidth connections or provide alternate versions of information products that require minimal data in order to download.

*Recommendations related to accessibility, low bandwidth access, and mobile access will be included in the OA Support Pack.*

---

<sup>11</sup>At the time of this writing, CG Core Metadata Schema is in development as a draft. The Data Management Taskforce will take ownership for CG Core after the taskforce has become operational.

## B: Storage, Copyright/Open Licenses, and Translations

Plans should address three elements related to the treatment of information products:

- Data storage and preservation (Article 4.1.4)
- Copyright and open licenses (Article 4.1.5)
- Translations (Article 4.1.7)

### B.1: Data Storage and Preservation for Future Use

Challenges associated with digital storage and preservation are twofold. First, a digital object must be preserved in such a way that the digital bits which comprise the object are able to be accessed in the future – i.e. ensuring the object itself is intact, that it is stored in such a way that its integrity is maintained. In this regard, it is recommended that Centers follow standard best practices for data storage and security such as redundancy and the use of SSL access to servers and systems.

The second aspect of storage and preservation is ensuring future usability of digital objects – i.e. the ability to interpret, understand, and use stored information. Digital preservation may require technologists to routinely migrate files from one format to another as formats become obsolete. Even so, technologists cannot guarantee future usability of proprietary formats.

As a result, in order to promote future use, Centers should encourage content creators to only deposit into repositories and other OA platforms information products which are in standard formats (e.g. PDF, Office Open XML, or plain text for documents; PDF, CSV, or XLS for data; MP3, MP4, MOV etc. for audio/video files, etc.)<sup>12</sup>

Centers might elect to offer different levels of preservation to different categories of information products based on uniqueness and importance.

At a minimum, plans should address:

- Steps taken to ensure secure storage of data such as storage redundancy;
- Acceptable or recommended file formats for storage of each type information product;
- Preservation mechanisms currently adopted;
- Steps taken to deal with file format obsolescence.

*Additional guidance from the Data Management Taskforce will be included in the OA Support Pack.*

### B.2: Copyright and Open Licenses

Article 4.1.5 states that *“Suitable open licenses shall be used that recognize the legal rights to information products and encourage their use and adaptation.”*

The license conditions upon which information products are made Open Access may vary depending on the nature of the information products and the need to limit or restrict access or usage rights to certain audiences and users. No single license is appropriate for all research projects.

---

<sup>12</sup>List of Open File Formats as maintained in Wikipedia: [http://en.wikipedia.org/wiki/Open\\_file\\_format](http://en.wikipedia.org/wiki/Open_file_format)

At a minimum, plans should:

- Offer Center-specific guidance or recommendations on adoption of open licenses;
- Provide recommendations for authors to ensure the originating Center maintains rights to translate key works;
- Identify which types of open licenses are commonly used, for what types of information products, and in what systems they are used;
- Identify ways in which re-use of information products is encouraged; and
- Identify a process to collect copyright and open licensing questions that arise in relation to the Open Access and Data Management Policy in order to share knowledge across Centers and inform development of Consortium-wide recommendations in this area.

*Additional guidance for researchers, particularly in terms of working with publishers and copyright agreements, will be issued as part of the Support Pack.*

### B.3: Translation

The Policy indicates in Article 4.1.7, *“Translations of key documents and other media into pertinent languages are encouraged. All versions should be deposited in suitable repositories and made Open Access.”*

At a minimum, plans should address:

- Tools embedded into repository and search functions to allow content to be discovered in and translated into relevant languages;
- Ways in which the Center has encouraged translations of key documents;
- Processes used to identify and track translations of key documents;
- Metrics used to monitor usage of translations of key documents through repository downloads; and
- Ways in which the Center is encouraging translations of key information products, particularly those which are targeted to reach specific beneficiaries.

*Additional guidance will be included in the Support Pack.*

## C: Administration of Open Access & Data Management

Implementing Open Access and data management will require time, infrastructure, financial resources and human expertise. Article 4.1.1 of the Policy states: *“Best efforts shall be used to make information products Open Access...”* ‘Best efforts’ is an intentionally high standard, which requires resources to be secured and expended in meeting the challenge of Open Access. Best efforts should include:

- Ensuring adequate staffing and expertise to support Open Access and data management
- Securing necessary financial resources

### C.1: Strategy and implementation

Support for Open Access and data management comes in many forms and levels, so no single model is appropriate for all Centers. It is recommended that Centers develop plans for Open Access and data management in a way that is in keeping with the Open Access and Data Management Policy and these guidelines and aligned with: current research agenda and priorities, technical infrastructure, staffing levels and areas of expertise, budget, and recent experience providing support for Open Access and data management.

While the Policy is designed to make final CGIAR Consortium information products available via Open Access, this transition process will take time. Article 2 of the Policy addresses this transition process: *“Implementation of and compliance with this Policy...will be phased over a transition period. The transition period runs from the effective date of the Policy for an initial period of 5 years, with comprehensive implementation by the end of 2018.”*

Further, not all CGIAR Consortium information products will be of equal value. Article 4.1.1 of the Policy states: *“Open Access arrangements should consider the characteristics of the information product, their potential impact, the level of data processing required, and whether the information products generated are within the scope of this Policy. Some judgment therefore needs to be made over the information products that will be made Open Access.”*

Centers – particularly those which do not have open access repositories in place or have not been making information products openly accessible in this manner – are encouraged to approach implementation as a staged process.

At a minimum, Centers’ plans should include a section related to strategy which addresses:

- Priorities for implementation
- Goals and objectives for the time period covered by the Policy
- Timelines and key milestones for the time period covered by the Policy

## C.2: Staffing

Open Access and data management support requires a mix of operational support and oversight/coordination. Furthermore, Open Access, data management, and the technical environment are constantly changing and will require ongoing professional development to support.

At a minimum, plans should address:

- Departments or individuals serving as the focal point for operational support for data management
- Departments or individuals serving as the focal point for operational support for Open Access
- A general workflow for depositing different types of information products into repositories
- Departments or individuals providing oversight or coordination of Open Access and data management
- Ways in which gaps in professional expertise will be addressed

*Additional guidance such as sample job descriptions for data managers will be issued in the OA Support Pack.*

## C.3: Financial Administration

Open Access and data management will require financial investments by Centers. Examples of anticipated or potential expenditures include:

- Article processing charges (APCs): charges levied by some Open Access journals and some “hybrid” closed-access journals that offer authors the ability to secure an open license for a particular article
- 
- Repositories: software (if using software that is not free/open source), hardware, server space, processing space

- 
- Staff: hiring new data managers, ongoing professional development for staff supporting Open Access and data management
- New tools: licensing of new tools to enhance discovery and encourage usage of Open Access content

Centers should begin to budget for Open Access and data management costs in future funding proposals, including the forthcoming second call for CRPs. As an interim step, the CGIAR Consortium is working towards making dedicated funds available to support Open Access and data management; it is intended that these funds will be allocated based on guidance from the OAIWG and DMTF. However, this is not a sustainable solution. Centers and researchers need to begin to budget for Open Access costs such as article processing fees in funding proposals. Alternatively, Centers may wish to consider alternate options such as recommending that authors publish in Open Access journals that do not levy article processing fees or determining what constitutes reasonable fees.

At a minimum, plans should address:

- 
- Budgets for infrastructure and appropriate staffing
- Budgets for professional development
- Fees for Open Access publishing
- Other major expected expenditures

*Additional guidance such as examples of mechanisms to support fee-based Open Access publishing will be forthcoming in the Support Pack.*

#### D: Assessment, Impact, and Review

The CGIAR Consortium will continue to work with experts across funders, private sector, academia, and society to develop meaningful metrics to assess usage, impact, and uptake for Open Access based on iterative learning and experimentation. New feedback, best practices, and overall expertise will be incorporated into the development of future Open Access objectives, milestones, and metrics. As such, it is expected that plans for assessing impact and uptake will evolve.

As a starting point, Centers' plans should address:

- Incentives (Article 4.1.6)
- Assessment and review of Open Access and data management at a programmatic level (Article 5)
- Tracking impact and uptake (Article 5)

Consortium-level plans for assessing and reviewing impact will be addressed in 2014-2015 as part of a second phase of implementation. During the first phase of implementation, Centers will report on their existing methods for incentivizing compliance.

##### D.1: Incentives

Since Open Access and data management is mandated through the Policy, all CGIAR Consortium researchers and other creators of information products are required to comply. Centers should encourage and reward sharing of information products (including data) through appropriate Open Access channels. Likewise, due consideration should be given in individual appraisals for compliance with the policy but also for taking extra steps to encourage discovery and uptake of information products.

At the minimum, plans should address:

- Ways in which researchers and authors/creators of information products are encouraged to comply with the Policy.
- Review processes or measure that reward scientists who are in compliance with the Policy.

Furthermore, Centers are encouraged to share any sample text from employment contracts or appraisal forms which reference (directly or indirectly) Open Access or data management.

*Specific examples will be collected and shared via the Support Pack.*

#### D.2: Assessment and review

Measuring successful implementation of the Policy will require both quantitative and qualitative information due to varying stages of information management maturity across CGIAR and the evolving nature of successful partner and public engagement. Centers are encouraged to develop their own methods for tracking progress of Open Access.

At the Consortium level, baseline information will be collected in 2014 and will be updated annually thereafter to track progress with implementation across Centers. Article 5 of the Policy states: “*The CGIAR Consortium Office will carry out an evidence-based review of the implementation of this Policy on an annual basis.*” The Policy itself will then be reviewed in 2015 and every two years thereafter.

At a minimum, plans should address mechanisms for internal review.

#### D.3: Tracking Impact and Uptake

As the scholarly communication landscape evolves through developments in Open Access and interoperable technologies, research institutions are developing and testing new ways to assess the impact of research outputs and track usage. “Altmetrics,” or new types of metrics which are only possible in the digital environment, are still experimental yet allow new ways to glean insight into how individuals interact with bodies of research or individual information products. While specific tools are still experimental, the scholarly community is embracing a shift towards new types of metrics to track usage in quantitative ways and trace the path of uptake in qualitative ways.

Centers are encouraged to experiment with new ways of measuring, assessing, and tracking research outputs. It is expected that once a cohesive infrastructure is in place, the Consortium will collectively test various tools and metrics to track usage and establish a common understanding for interpreting these new metrics.

At the minimum, plans should address:

- Which metrics are being collected, how these metrics are collected, and how they are interpreted in order to understand usage, impact, and uptake of materials disseminated through Open Access

*Additional resources will be forthcoming through the Support Pack to help address altmetrics and other emerging ways to track impact and uptake.*

### 3. CGIAR Open Access-Open Data Implementation Plan Template

- *Legal*
  - *Access Rights*
  - *Licensing of data and research*
  - *Open access/restrictions*
  - *Ownership of data*
  - *Partnership agreements*
- *Managerial*
  - *Budgeting*
- *Strategic Planning*
  - *Data Ownership*
  - *Project level Data Management and Sharing Strategy/Plan*
  - *Engaging and interacting with project partners*
  - *Managing expectations and rights of partners*
- *Technical*
  - *Data security*
- *Reference*
- *When*
  - *Decisions while designing*
  - *Management of research processes*
  - *Delivery of research products*
- *PI*

Please find below the document



---

# Open Access/Open Data Implementation Plan

<MEMBER CENTER NAME>

---

Version <2.0>

<Date>

*[Notes to Open Access and Open Data Team]*

*[This document is a template of an Open Access/Open Data Implementation Plan document which complies with the criteria specified by the CGIAR Open Access and Data Management Policy and the CGIAR Open Access and Data Management Implementation Guidelines. The template includes instructions to authors, boilerplate text, and fields that should be replaced with the values specific to the Center/CRP.]*

- Blue italicized text enclosed in square brackets ([text]) provides instructions to the document authors or describes the intent and context for content included in the document.*
- Blue italicized text enclosed in angle brackets (<text>) indicates a field that should be replaced with information specific to a particular Center/CRP.*
- Text and tables in black are provided as boilerplate examples of wording and formats that may be used or modified as appropriate. These are offered only as suggestions to assist in developing implementation plans; they are not mandatory fields, formats, or text.]*

*[When using this template, the following steps are recommended:*

- 1. Replace all text enclosed in angle brackets (e.g. <Center Name>) with the correct field/document values.*
- 2. Modify boilerplate text as appropriate.*
- 3. To add any new sections, ensure that the appropriate header and body text are used in order to maintain styles. Styles used for this document are: Heading 1, Heading 2, and Heading 3. Style used for boilerplate text is Normal.*
- 4. Before finalizing the plan, delete this section and all author instructions.*
- 5. To update the Table of Contents, right-click on it and select "Update field" and choose "Update entire table."]*

*[Please note: If your Center/CRP already has an Implementation Plan (or one in development), it is not necessary to revise the plan using this template. However, please consider adding the information and detail included in the template to the Center or CRP's plan.]*

## Section 1: Introduction

### 1.1: Purpose of the OA/OD Implementation Plan

*[This subsection presents a very brief introduction to <Center/CRP's> implementation of Open Access and Open Data.]*

Open Access is vitally important to increasing the visibility, accessibility and impact of the research of CGIAR and other agricultural research for development stakeholders. A number of reports and studies have established that Open Access not only increases the visibility of research but also increases the citation rate of research and the utility of underlying data. Furthermore, a rigorous and consistent approach to data management will ensure that data is collected, stored, analyzed and shared in a manner that will have the greatest impact, whilst also protecting the rights of third parties and stakeholders where appropriate.

*[Note: If the Plan has been written by a team that includes external authors, please include a short footnote identifying these authors and offering a brief statement about their expertise in this arena.]*

### 1.2: Scope of Open Access and Definition of Openness

*[This subsection includes a description of the scope of Open Access applicable pursuant to the CGIAR Open Access and Data Management Policy.]*

This OA/OD Implementation Plan has been developed pursuant to the CGIAR Open Access and Data Management Policy (adopted 2013) and the CGIAR Open Access and Data Management Implementation Guidelines (adopted 2014). This policy framework stipulates that Open Access is required to all CGIAR information products, with the exception of those subject to narrow limitations such as:

- a) Final information products produced prior to 2 October 2013 (i.e. the effective date of the CGIAR Open Access and Data Management Policy);
- b) Information products that are unstable, unlikely to undergo further change or contain characteristics which are assessed to be of limited value to others (e.g. due to low quality);
- c) Information that is determined to be of a sensitive nature due to considerations including privacy, price and political sensitivity, adverse effects on farmers rights, etc.
- d) Confidential information associated with permitted restrictions or subject to limited delays to seek IP rights pursuant to the CGIAR IA Principles;
- e) Confidential information of Centers beyond the scope of the CGIAR Open Access and Data Management Policy or the IA Principles (for instance, HR hiring documents, personnel records, certain types of financial records, certain types of contracts or vendor agreements, private Board of Trustee minutes all include sensitive and/or confidential information and will not be included in OA repositories).

*[This subsection includes a definition of openness – either the definition included in the CGIAR Open Access and Data Management Policy or the Center’s definition if it differs.]*

For the purposes of the CGIAR Open Access and Data Management Policy, Open Access means the immediate, irrevocable, unrestricted and free online access by any user worldwide to information products, and unrestricted re-use of content (which could be restricted to non-commercial use and/or granted subject to appropriate licenses in line with the CGIAR IA Principles), subject to proper attribution.

*[Please note that while the CGIAR definition of open access permits restrictions as to commercial use, increasingly, research funding agencies such as the Bill & Melinda Gates Foundation require unrestricted reuse of content, including for commercial purposes, by mandating use of a CC-BY 4.0 International License or equivalent) to be considered “open.”]*

### 1.3: Overview of the <Center/CRP>

*[A brief – approximately 1 paragraph – introduction to the Center/CRP with regard to the size, scale, and scope of Open Access and Open Data operations to offer context for readers.]*

*[Examples of items to note: approximate number of FTE researchers, annual research budget, key funding agencies with OA/OD policies that will likely affect this Center/CRP’s researchers, approximate number of peer-reviewed articles published each year, and journals where researchers from this Center/CRP publish on a regular basis.]*

### 1.4: Overview of Current OA/OD Environment at <Center/CRP>

*[This subsection briefly describes the current Open Access/Open Data environment at the Center and which open repositories are in use.]*

*[If a Center/CRP has its own OA/OD Policy, it should be noted here. Furthermore, please indicate how the policy differs from the CGIAR Open Access and Data Management Policy.]*

### 1.5: Information Products and Priorities

*[This subsection indicates which types of information products are the focus of the implementation plan. It is expected that some Centers might wish to prioritize certain types of information products or content within product types during the initial phase of implementation. If so, this should be noted, along with the framework used for prioritization (e.g. of particular datasets). It is recommended that peer-reviewed, scholarly articles published after 02 October 2013 and datasets for projects which were completed after 02 November 2013 should be the top priority during this initial phase of implementation.]*

*[Specify which types of information products are prioritized for Open Access treatment.]*

### 1.6: Deposit Schedules for Information Products

*[The Open Access & Data Management Implementation Guidelines indicate that “content should be deposited in full as soon as possible after an item is complete or is in its final form. Plans should address timelines for depositing information products into repositories. Plans’ deposit schedules should be consistent with these guidelines or indicate shorter timelines than those presented here. See Table 1 for details from the Open Access & Data Management Implementation Guidelines.]*

*[If the implementation plan allows for a longer gap between the completion of an information product and its deposit into a repository during the transition phase, include an explanatory note.]*

The timeframes stated in the Implementation Guidelines (Table 1) reflect the minimum deposit commitments made by CGIAR Centers during the transition period until 2 October 2018, after which the deposit schedule contained in the Policy becomes binding.

**Table 1: Deposit Schedules from the CGIAR Open Access & Data Management Policy and Implementation Guidelines**

<b>Types of Information Products</b>	<b>Transition Deposit Schedule (until October 1, 2018)</b>	<b>Policy Deposit Schedule (from October 2, 2018)</b>
Peer-reviewed versions of journal articles	As per the Policy Deposit Schedule unless OA is prohibited or subject to a longer embargo period by publisher	Ideally, at the time of publication Latest: 6 months from publication
Self-published journals, books, reports etc.	Immediately	Self-published materials not currently addressed in the Policy
Reports and other papers	As soon as possible Latest: within 6 months of completion	As soon as possible Latest: within 3 months of completion
Externally or commercially published books and book chapters	As per the Policy Deposit Schedule	As soon as possible Latest: within 6 months of completion
Data and data sets	As per the Policy Deposit Schedule	As soon as possible Latest: within 12 months of completion of data collection or appropriate project milestone, or within 6 months of publication of the information products underpinned by that data
Video, audio, scientific images	As soon as possible Latest: within 6 months of completion	As soon as possible Latest: within 3 months of completion
Photographs	As soon as possible Latest: within 6 months of completion or publication	As soon as possible Latest: within 3 months of completion or publication
Computer software/applications/code	As soon as possible Latest: within 6 months of completion	Upon completion of software development
Metadata	As soon as possible Latest: before or on publication of the information product	As soon as possible Latest: before or on publication of the information product
Core/corporate governance documents appropriate for public consumption	e.g., financial reports, board agendas and minutes, annual reports, as appropriate As soon as possible	As per 'reports' category of Information Product (Core/corporate governance documents not currently addressed separately in the Policy)
Automated deposit extensions	Certain types of information products (in particular data collected pursuant to hypothesis-driven research) may take longer than 12 months to clean, analyze and publish. Thus, 12 months should be seen as the aim, with 24 months as the long-stop date for making such data Open Access.	A long-stop date of 24 months is not currently included in the Policy

While the Policy is designed to make final information products available via Open Access as quickly as possible, this transition process will take time; as such, implementation may occur through a phased approach. The Implementation Guidelines are intentionally broad in nature and designed to offer Centers/CRPs as much flexibility as possible while they plan for and prepare their own implementation plans and approaches to supporting Open Access and Data Management. There is, therefore, allowance for a slightly longer gap between completion of research and deadlines by which information products will be expected to be deposited into repositories.

*[If the Center/CRP is adopting different timelines from the above table in regard to the transition period, it should be noted here and the rationale explained.]*

### 1.7: Exceptions and Extensions to the Deposit Schedules

*[Centers/CRP need to explain the internal notification and approval mechanisms through which exceptions and non-automated exemptions to the deposit schedule will be managed with regard to (a) information that is determined to be of a highly sensitive nature due to considerations including privacy, price and political sensitivity, adverse effects on farmers rights, etc.; and (b) confidential information as may be associated with permitted restrictions or subject to limited delays to seek IP rights as per the CGIAR IA Principles. In both scenarios the involvement of the Data Manager and/or Knowledge Manager as well as the IP Focal Point<sup>1</sup> is strongly recommended.]*

## Section 2: Strategy and Implementation Overview

### 2.1: Overview of Strategy and Approach to Implementation

*[Overview of the approach the Center/CRP is taking in its implementation of Open Access and Open Data – i.e., if the Center/CRP is approaching implementation for both areas as a single work stream, or if the Center/CRP is handling Open Access to publications via one stream and Open Data separately.]*

### 2.2: Goals and Objectives

*[Center/CRP-specific goals and objectives for Open Access and Open Data.]*

*[From the CGIAR Open Access and Data Management Policy: CGIAR regards the results of its research and development activities as international public goods and is committed to their widespread dissemination and use to achieve the maximum impact to advantage the poor, especially smallholder farmers in developing countries. CGIAR considers Open Access (defined below) to be an important practical application of this commitment as it enhances the visibility, accessibility and impact of its research and development activities. Open Access improves the speed, efficiency and efficacy of research; it enables interdisciplinary research; assists novel computation of the research literature; and allows the global public to benefit from CGIAR research. Furthermore, CGIAR recognizes the benefits that accrue to individual researchers and to the research enterprise from wide dissemination, including greater recognition, more thorough review, consideration and critique, and a general increase in scientific, scholarly and critical knowledge. CGIAR further recognizes that, in*

---

<sup>1</sup> Involvement of the IP Focal Point is recommended because the application of an exemption or extension to the deposit schedule has implications concerning a Center's obligations to promptly and broadly disseminate results as required pursuant to Article 6 of the IA Principles.

*implementing this Policy, it can more easily and collectively build the infrastructure necessary to be at the forefront of the open access and open data for agriculture movement.]*

*[If the goal is to increase uptake and usage of CGIAR research and development activities, objectives for OA/OD should be more concrete. For instance, to have 75% of research produced in 2014 fully openly-accessible via the Center's repository by 1 July 2015.]*

### 2.3: Timelines and Key Milestones

*[This subsection should outline activities and milestones in detail for 2015 and at a high level for 2016-2018. If a Center/CRP is pursuing a phased approach to implementation, it should be noted and explained in this subsection.]*

### 2.4: Anticipated Needs and Challenges

*[This subsection should identify any significant anticipated challenges for implementing OA/OD within the Center/CRP. Some items to consider:*

- *Schedule*
- *Budget*
- *Resource availability and skill sets*
- *Technology to be acquired, updated, maintained, installed, etc.*
- *Researchers' awareness, compliance, challenges*
- *Administering article processing charges (APCs) at scale*
- *Supporting researchers' questions and concerns around OA publishing and deposits*
- *Supporting researchers' questions and concerns around data management and data quality]*

### 2.5: Lead Centers, Participating Centers and Partners

*[Use this subsection to indicate the relationship between the Center and CRPs – for example, which CRPs this Center is serving as the lead Center and which CRPs this Center is serving as a participating Center. The assumption is that a Center's Implementation Plan will be applicable to the CRPs it leads.]*

Information products produced by lead Centers and participating Centers (including partners) in CRPs are subject to the Open Access and Data Management Policy. Agreements put in place after 2 October 2013 should be carefully negotiated to ensure that any restrictions on sharing data under a research and/or development project are limited in duration, territory and/or field of use, if applicable, and fully justifiable by reference to the CGIAR IA Principles (i.e., in particular, articles 6.2, 6.3, and 6.4)<sup>2</sup> and the CGIAR Open Access and Data Management Guidelines.

## Section 3: Technical Infrastructure

*[Many Centers/CRPs use separate repositories for data and publications. In these cases, it might be preferable to split this section into two – one section for the Technical Infrastructure for Open Access Publication Repositories and a second section, including the same elements, for Technical Infrastructure for Open Data.]*

---

<sup>2</sup> Article 6.2 is on "Limited Exclusivity Agreements," Article 6.3 is on "Incorporation of Third Party Intellectual Assets" and Article 6.4 is on "Intellectual Property Rights," all part of the CGIAR IA Principles.

### 3.1: Repository Systems

*[Repository systems should meet current industry standards for interoperability and metadata. Plans should address which repositories are in use for each type of information product and associated URL(s).]*

*[For Publication and General-Purpose Repositories: DSpace, EPrints, Invenio, and ContentDM are all acceptable repositories and meet the required standards for interoperability and metadata.]*

*[For Data Repositories: Dataverse meets the required standards for interoperability and is currently in use by many CGIAR Centers to collect and disseminate datasets. During 2015, the DMTF is expected to issue recommendations for other types of data and GIS repositories.]*

*[If a Center/CRP is using a system other than the ones listed above, it must be able to allow for the transfer of metadata via a protocol such as OAI-PMH or OData, and the digital objects must be openly accessible per the guidelines set forth in the CGIAR Open Access and Data Management Policy. Specifically, the digital objects themselves (postprints or publishers' versions of articles, complete data sets) must be freely and directly accessible to the public.]*

### 3.2: Interoperability

*[This subsection should identify which interoperability protocols, standards, APIs, web-based aggregation, and harvesting systems, services, and tools are in use.<sup>3</sup> In addition, please address any current or planned ways in which digital objects and/or metadata will be transferred or interpreted between systems (internally or externally), and any new or emerging tools, protocols, or frameworks being tested or considered for adoption.]*

*[At a minimum, all repositories address:*

- *Which mechanisms are in place to enable cross-system transfer of content or metadata and how these mechanisms are being used – for instance, if SWORD or Dataverse APIs are in use;*
- *How content is transferred between systems (internally or externally);*
- *How metadata is transferred between systems (internally or externally);*
- *Whether each repository is OAI-PMH compliant and if OAI-PMH is currently enabled*
- *What other interoperable tools are enabled and adopted within the repositories – for instance, SWORD, LOD, Dataverse APIs, Altmetrics or any other relevant interoperability tools/plugins*
- *Any new or emerging tools, protocols, or frameworks under consideration for adoption in the next 1-2 years]*

### 3.3: Metadata

*[Centers/CRPs should demonstrate that they are working towards adopting the CG Core Metadata Schema. Plans should confirm that they have adopted CG Core and offer details on any additional metadata schemas and taxonomies are in use and how they are applied. Please include a crosswalk to demonstrate how repository fields are related to CG Core elements.]*

---

<sup>3</sup> For more details about interoperability and repositories, please consult the following reports produced by the Confederation of Open Access Repositories (COAR): “The Case for Interoperability for Open Access Repositories,” “The Current State of Open Access Repository Interoperability (2012),” and “The COAR Roadmap: Future Directions for Repository Interoperability.” All three reports are accessible at: <https://www.coar-repositories.org/activities/repository-interoperability/>



*[In the long run, Centers/CRPs should aim to adopt Linked Open Data protocols as appropriate such as the AGROVOC Linked Open Data API to use AGROVOC terms from within a repository. Include any details regarding plans for moving towards implementing Linked Open Data.]*

*[CG Core refers to use of persistent identifiers. Please indicate which type(s) of persistent identifiers are implemented and in which repository.]*

### 3.4: Data Storage and Preservation for Future Use

*[Plans should address storage and preservation issues such as storage redundancy; storage formats and preservation mechanisms; use of persistent identifiers; recommended and accepted file formats.]*

*[Plans – and repositories – should include a preservation policy. For example:]*

<Center/CRP> is committed to responsible and suitable management of works deposited in <repository name>.

1. Digital preservation is an evolving field. <Center/CRP> bases its preservation strategy on the Open Archival Information System (OAIS) reference model (ISO 14721:2012); this strategy will continue to evolve and is informed by current and emerging best practices.
2. Efforts will be taken to preserve any work submitted to <repository>. However, contributors are strongly encouraged to deposit information products in a recommended file format in order to facilitate long-term preservation. See <url/list of recommended file formats> for details. For files in other formats, a derivative copy in a more stable format should be created if feasible. In these cases, both versions and associated metadata should be deposited.
3. <Repository> will provide long-term access to submitted works along with associated metadata. In order to provide long-term access, <Center/CRP> will: backup files in a secure and redundant manner, periodically refresh the storage media, and migrate obsolete file formats for files stored in recommended open file formats.
4. At this time, <Center/CRP> is committed to preserving the bitstream of files.
5. All works submitted to <repository> will receive a persistent URL.
6. This policy will be reviewed annually to ensure practices are consistent as technology and best practices evolve.

*[Centers/CRPs are encouraged to maintain a list of recommended file formats on the repository website, but please include a current list as part of the implementation plan or a url to a live website.]*

Recommended file formats for publications:

Format	File Extensions
Acrobat PDF/A	.pdf
Comma-separated values	.csv
Open Office formats	.odt, .ods, .odp
Plain text (US-ASCII, UTF-8)	.txt
XML	.xml

*[If images, audio, and video files are collected and disseminated via the repository, recommended formats should be indicated as well. Ex: [http://www.lib.cam.ac.uk/dataman/resources/File\\_Formats.pdf](http://www.lib.cam.ac.uk/dataman/resources/File_Formats.pdf)]*  
*[Data repositories should include recommendations for preferred file formats.]*

*[Perpetual Access: Plans should include a statement indicating what happens if the repository is shut down.]*

### 3.5: Limited Internet Connectivity

*[This subsection should include ways in which the Center/CRP is specifically supporting access to content in low-bandwidth and mobile environments.]*

To assist those with limited internet connectivity, designing easily accessible information products or providing alternate versions of materials that require minimal data download is encouraged if it does not affect the quality of the information products.

In order to maximize uptake within these environments, <Center/CRP> has taken the following steps, based on Aptivate’s best practices:<sup>4</sup>

- Whenever feasible, pages are smaller than 25kB
- <etc.>

## Section 4: IPR/Intellectual Assets

### 4.1: CGIAR Principles on the Management of Intellectual Assets

*[This subsection includes a brief statement connection Open Access to the CGIAR IA Principles.]*

The CGIAR IA Principles and associated implementation guidelines provide for the prompt and broad dissemination of research results, which translates to a default assumption that information products should be made openly accessible as soon as possible. Article 1 of the Open Access and Data Management Policy states “this Policy complies with the CGIAR Principles on the Management of Intellectual Assets (‘CGIAR IA Principles’), which is the umbrella document for this Policy.” The assumption of Open Access may be challenged by reference to the allowable/reportable restrictions and exclusions set out in the CGIAR IA Principles and associated implementation guidelines, and the guidance in these Guidelines that allow for restricting Open Access in certain circumstances.

### 4.2: Open Licenses

*[Center-specific guidance related to suitable open licenses.]*

Article 4.1.5 states that ‘*Suitable open licenses shall be used that recognize the legal rights to information products and encourage their use and adaptation.*’

In addition to providing greater access to knowledge, Open Access and Open Data also include provisions for allowing for the rights for others to reuse information products – which means using appropriate open licenses.

---

<sup>4</sup> Web Design Guidelines for Low Bandwidth by Aptivate: <http://www.aptivate.org/webguidelines/TopTen.html>

*[A variety of open licenses exist; Centers/CRPs should work with their IP focal points to recommend specific licenses via the Implementation Plan. For publications and data, Creative Commons Attribution licenses (CC-BY 4.0) are becoming the norm for Open Access and Open Data policies (and is in fact mandated by the Bill and Melinda Gates Foundation).<sup>5</sup> The GNU General Public License (GNU GPL) is often used for software and programming code.<sup>6</sup>]*

#### 4.3: Guidance for Authors

*[It is recommended that Centers include or link to guidance for authors in their implementation plans.]*

#### 4.4: Translations

*[Use this subsection to address how recommended licenses will allow for re-use such as translations of information products and ways in which translations are encouraged. If specific open licenses were recommended in section 4.2, it is important that these licenses are open enough to allow for translations and other types of re-use.]*

Article 4.1.7 of the CGIAR Open Access and Data Management Policy states that ‘Translations of key documents and other media into pertinent languages are encouraged. All versions should be deposited in suitable repositories and made Open Access.’

<Center/CRP> encourages adoption of CC-BY licenses; these licenses allow for re-use of information products, including translations.

### Section 5: OA/DM Teams and Staffing

*[Use this section to describe who is responsible for Open Access and Open Data operations at the day-to-day operational level, management and oversight of Open Access and Open Data, and the make-up of any cross-functional teams. Reconfigure this section as appropriate based on the Center’s organizational structure.]*

#### 5.1: Day-to-Day Operations

*[Describe which team members and/or individuals are responsible for day-to-day operations of Open Access and Open Data/Data Management.]*

#### 5.2: OAIWG and DMTF Representation

*[Indicate who the Center/CRP’s representatives are for the CGIAR-wide Open Access Implementation Working Group (OAIWG) and Data Management Task Force (DMTF).]*

#### 5.3: <Center/CRP Steering Committee> and Other Internal Partners

*[Use this subsection to describe any Center-specific oversight teams, steering committees, and internal partners – for example:]*

<Center> has an Open Access group with representation from <the library, KM team, Communications Department, IT, and Legal>. The group is chaired by <name> from <department/team>. The full team includes:

---

<sup>5</sup> Creative Commons Attribution 4.0 License: <http://creativecommons.org/licenses/by/4.0/>

<sup>6</sup> GNU General Public License: <http://www.gnu.org/licenses/#GPL>

- <name>, <department/team>
- <name>, <department/team>
- <name>, <department/team>
- <name>, <department/team>

## Section 6: Promoting and Supporting Researchers' Implementation of OA for Publications

*[Most of the information included in this section is not explicitly referenced in the Open Access and Data Management Policy or implementation guidelines; however, Centers/CRPs are highly encouraged to address these issues as part of the implementation planning process, and thus it is recommended that these subsections are included in implementation plans. Furthermore, the information captured in these sections will be useful for other CGIAR Centers to generate ideas and share best practices.]*

### 6.1: Deposit Workflows for OA Publications Repository

*[This subsection should be used to include a high-level overview of the process to deposit an item into the publication repository. Describe in words or with a simple diagram, whichever is easier.]*

*[If the Center/CRP does not yet have an operational repository, it is important to indicate how materials are being captured, collected, and described as an interim step while the repository is being developed.]*

### 6.2: Author Guidance

*[This subsection should capture and share current plans for dealing with issues specific to OA for publication such as guidance for authors, using author addendum, information about Creative Commons licenses, guidance on where to publish, etc. If such materials are available as standalone documents, please include urls. If such items are only available within a locked intranet, please attach as an annex.]*

### 6.3: Funding for OA Fees

*[This subsection should capture and share current plans for dealing with issues specific to expenses related to Open Access publishing fees – article-processing charges (APCs), hybrid OA journal fees (i.e. payments to allow authors to opt-in to publish with an OA license or gain permission to deposit in an OA repository), guidance for integrating OA into funding proposals, etc. If such materials are available as standalone documents, please include urls. If such items are only available within a locked intranet, please attach as an annex.]*

*[If the Center/CRP has negotiated any Center-specific fees or arrangements with publishers, please explain here.]*

### 6.4: Internal Communication Strategy

*[Overview of communication strategy and ways to raise awareness about implementation among researchers. The internal communication strategy or advocacy plan should be approached jointly for Open Access and Open Data in order to present a cohesive message.]*

---

## Section 7: Promoting and Supporting Researchers' Implementation of Open Data and Data Management

### 7.1: Deposit Workflow for Open Data Repository

*[This subsection should be used to include a high-level overview of the process to deposit an item into the data repository. Describe in words or with a simple diagram, whichever is easier.]*

*[If the Center/CRP does not yet have an operational repository, it is important to indicate how materials are being captured, collected, and described as an interim step while the data repository is being developed.]*

### 7.2: Support for Data Management Practices and Data Quality

*[This subsection should capture and share current plans for dealing with issues specific to Open Data such as when to publish data in a repository or working with data citations. If such materials are available as standalone documents, please include urls. If such items are only available within a locked intranet, please attach as an annex. Please use this section to highlight how the Center/CRP is dealing with data quality.]*

### 7.3: Data Streams

*[CGIAR research covers a wide range of subject areas, sectors, and data streams, including but not limited to, data related to: genetics/genomics; genebanks; agronomy; breeding; natural resource management – including soils, hydrology, climate, and more; socioeconomics – including surveys, food security, poverty, livelihoods, nutrition, and allied areas; geospatial information, and other sectors.]*

*[Several of these data streams might require subject-specific support – for instance, different types of tools, services, systems, or metadata to further enhance discovery and usability of data from that stream and/or allied areas. Please use this section to highlight how the Center/CRP is dealing with nuances related to particular data streams –ways in which a data stream is being supported in a different way or through any specialized services, tools, systems, or metadata.]*

## Section 8: Financial Administration

### 8.1: Major Expenses

*[Use this section to present the budget for Open Access/Open Data. Identify any significant expenses.]*

**Table 2: Open Access/Open Data Budget for 2015**

Line Item	Amount	Explanatory Notes
<b>IT/Infrastructure:</b>		
Data Repository		
Publications Repository		
Hardware		<Computers, servers, other equipment>
Programming/development		<identify which repository>

Annual maintenance fees		<identify which repository>
Website development related to repositories		<identify which repository/website>
<Other>		<Ex: DOIs or other types of persistent identifiers>
<b>Staffing:</b>		
Staff salaries – open data		<Indicate approx. FTE for Open Data>
Staff salaries – OA publications		<Indicate approx. FTE for Open Access for Pubs>
Professional development for OA/DM		
<b>Membership Fees:</b>		
Altmetrics provider(s)		<Indicate which companies/services>
Publisher-based institutional memberships		<Indicate which – ex: PLOS Institutional Account, Springer OA Membership>
<Other>		<Indicate which – ex: DataCite membership, RDA membership, COAR membership>
<b>Other Expenses:</b>		
Marketing/promotion materials		
OA Fees for Articles		<Article-processing charges – total if the Center is paying for these fees vs. authors incorporating into grant funding or waived fees>
<Other>		

*[In addition to the budget for 2015, please indicate any major upcoming expenses for 2016-2018 -- for example, development work leading up to a system migration or costs associated with starting up a new repository.]*

## Section 9: Assessment, Impact, Review

*[This section should be used to address any Center-specific plans for tracking and assessing impact of Open Access and Open Data and their measures of success. It is recommended that Centers/CRPs review progress, compliance, and uptake at multiple levels: at the highest level, by looking at the repository itself; gauging usage and/or uptake for content within repositories (item/article level); and examining individuals' compliance. Furthermore, please indicate any specific steps in place to maximize visibility of the Open Access/Open Data initiative, repositories, and repository content.]*

*[Plans should address which metrics are being collected and how they are interpreted in order to understand usage, impact, and uptake of materials disseminated via open repositories. Input from these sections will inform system-level recommendations for assessment. Further recommendations related to metrics, altmetrics, quantitative and qualitative measurements are expected later in 2015 based on input from Centers, CRPs, the OAIWG, and DMTF.]*

### 9.1: <Center/CRP> Repository-Level Metrics

*[Management of OA/OD efforts should include repository-level metrics to track and measure growth of repository content, who is adding records, how much repository items are downloaded, country-level information about who is making downloads, and how users are finding repositories.]*

## 9.2: Measuring Item-Level Usage/Uptake

*[This subsection should include statements about altmetrics, article-level metrics, data citations, and metrics for data. Indicate any altmetrics providers the Center/CRP is using such as Plum Analytics, Impact Story, Altmetric.com. Also how those altmetrics are being used and how they are shared with researchers.]*

*[Include a statement about Impact Factor – if it is being used for any purpose, how so.]*

## 9.3: Measuring Individuals' Compliance

*[Plans should address ways in which individuals' compliance is tracked, measured, and reviewed. If the Center/CRP has mechanisms in place to push compliance, it should be noted here. For instance, the University of Liege in Belgium only recognizes those items deposited into repositories as part of the annual review process.]*

*[Some research-funding organizations like the U.S. National Institutes of Health (NIH) policy and the Wellcome Trust are beginning to penalize researchers who do not comply with OA/OD policies, resulting in withholding or delayed payment of funds.<sup>7</sup>]*

*[Employment contracts should be updated to include reference to the OA/OD policy and appropriate methods of assessment.]*

Credit should be given to researchers for efforts to disseminate their information products in openly-accessible ways (i.e. via depositing into the <Center's/CRP's> repositories). Employment contracts now include the following language to include reference to the OA/OD Policy and how it will be tied to assessment:

*[Insert language here.]*

## 9.4 Assessing and Reviewing <Center/CRP>-Level Progress and Impact

*[This subsection should indicate how progress at the Center/CRP-level is being tracked and how it will be interpreted. Some metrics to consider: % of full-text articles vs. metadata-only records; the number of full-text articles submitted within 1 month of publication, within 3 months of publication, within 6 months of publication; etc.]*

## 9.5: Increasing Visibility – Additional Steps

*[Use this section to describe any other ways in which the repository, repository contents, and CGIAR Open Access/Open Data are being promoted in order to increase visibility and uptake of research. For example, list any directories and registries in which the repository is listed, any known harvesters that are harvesting and aggregating contents, suggestions offered to researchers on increasing visibility of their materials, etc.]*

---

<sup>7</sup> For further details on non-compliance, see Richard Van Noorden, "Fundlers punish open-access dodgers," in *Nature News* (09 April 2014): <http://www.nature.com/news/funders-punish-open-access-dodgers-1.15007>

---

# Open Access/Open Data Implementation Plan

## Template Annex

---

**11 May 2015**

*About the Template Annex:*

Additional suggested text and tables that were requested during the Asia and MENA Open Access/Open Data Implementation Workshops are included in this document.

The notes in this Annex are intended to provide supplemental guidance, clarifications, or revised tables to include in Centers'/CRPs' Implementation Plans based on the feedback and input received from participants during these two workshops.

Materials included in this annex are not intended to replace any content in the Template, but rather to offer additional language to consider incorporating into implementation plans.

Change log:

- Section 1.1 – Additional text added May 2015
- Section 1.7 – Additional text added May 2015
- Section 8.1 – Additional text added May 2015



## Section 1: Introduction

### 1.1: Purpose of the OA/OD Implementation Plan

*[This subsection presents a brief introduction to <Center/CRP> implementation of Open Access and Open Data. Some Centers requested additional sample text to incorporate referring to Open Access policies by key donors.]*

Over the past five years, major research funders around the world have begun to mandate Open Access to outputs of the research they fund. In late 2014, the Bill and Melinda Gates Foundation instituted an important policy which, after a two-year transition period, will require immediate, unrestricted access and reuse to all research outputs – including data – funded in part or in whole by the foundation. Other major funders of CGIAR research such as DFID, USAID, and USDA also have their own policies which push for open access to research outputs, including data.

### 1.7: Exceptions and Exemptions to the Deposit Schedule

*[Several Centers requested further clarification and sample text to use as a starting point for this section. Further clarification and a suggested work flow is provided below. The workflow should be revised based on the actual workflow in place at each Center.]*

*CGIAR Centers have a general legal obligation to promptly and broadly disseminate research results (as per Article 6, CGIAR IA Principles) and specifically, to make information products Open Access in accordance with the deposit schedules contained in the Open Access Implementation Guidelines (covering the transition period until October 2018) and the Open Access Policy (binding as from October 2018). To facilitate legal compliance and to minimize reputational risk to the CGIAR as a whole that can arise from instances of non-compliance, all requests for exceptions and/or exemptions to the Policy should be made in writing and follow a clear internal procedure for review on a case-by-case basis. Additionally, good faith compliance with the Open Access Policy during the transition period should ensure internal tracking of instances in which the deposit schedule timeframes specified in the Implementation Guidelines cannot be achieved. This includes instances in which the long-stop date of 24 months is relied upon as an automated deposit extension, as well as exceptional instances during the transition period in which despite best efforts, the long-stop date of 24 months cannot be achieved. This evidence collection will help inform evaluation during the transition period and any future review of the Open Access Policy and its Implementation Guidelines.]*

Any determination that research outputs (i.e., publications, data, or other types of information products resulting from research) should not be made openly accessible because it is subject to confidentiality or is of a highly sensitive nature, should be submitted in writing for approval by <Deputy Director General> and <IP Focal Point/Legal Department>. For instance, exceptions to the policy may be made for materials that might adversely affect farmers' rights, have privacy implications for individuals, are politically sensitive in nature, or contain pricing details that could negatively impact farmers.

For instances in which the deposit timeframes specified in Section 1.5 cannot be achieved, researchers should notify in writing the <Data Manager, KM Focal Point, or appropriate member of OA/DM team>. Notification should be sent upon the lapsing of the deposit schedule or sooner if known in advance, and should include an explanation of the challenges which have delayed or prevented deposit as well as any other relevant information to facilitate internal follow-up and assistance. Likewise, when the long-stop date of 24 months cannot be achieved, researchers should notify in writing the <IP Focal Point/Legal Department> **and** the <Data Manager, KM Focal Point, or appropriate member of OA/DM team>. The notice is to be given upon the lapsing of the 24 months or earlier if delays are anticipated

in advance. The notice should include an explanation of challenges experienced which have delayed or prevented the deposit and any other pertinent information which may facilitate internal follow up and assistance. The information collected in these notices will serve as evidence to inform evaluation during the transition period and any future review of the Open Access Policy and Implementation Guidelines.

## Section 8: Financial Administration

### 8.1: Major Expenses

*[Researchers are encouraged to incorporate the cost of dissemination into project proposals and funding requests. Likewise, CGIAR CRPs should include the cost of research dissemination into their planning efforts, particularly as part of the 2<sup>nd</sup> round of CRP requests. However, for the next few years, we will be in a transition period while research that has already been funded and did not include dissemination costs or data management into projects still needs to be made openly accessible.]*

*[In order to separate initial costs and one-time fees to operationalize Open Access, particularly during this transition period, Centers/CRPs are encouraged to consider their budgets for the transition period and also after the second round of CRPs are put into effect. A sample table is provided below. One-time costs should be noted – for instance, repository start-up fees, content migration into suitable repositories, or one-time metadata enhancements to align with CG Core.]*

Line Item	Annual amount 2015 – 2017 (transition period)	Annual amount 2018+ (after 2 <sup>nd</sup> round of CRPs in effect)	Explanatory Notes
<b>Technology</b>			
Data Repository			
Publications Repository			
Hardware			<Computers, servers, etc.>
Programming/development			<identify which repository>
Annual maintenance fees			<identify which repository>
Website development related to repositories			<identify which repository/website>
<Other>			<Ex: DOIs or other types of persistent identifiers>
<b>Staffing</b>			
Staff salaries – open data			<Indicate approx. FTE for Open Data>
Staff salaries – OA publications			<Indicate approx. FTE for Open Access for Pubs>
Professional development for OA/DM			
<b>Membership Fees</b>			
Altmetrics provider(s)			<Indicate companies/ services>
Publisher-based institutional memberships			<Indicate which – ex: PLOS Institutional Account, Springer OA Membership>
<Other>			<Indicate which – ex: DataCite membership>
<b>Other Expenses</b>			

Marketing/promotion materials			
OA Fees for Articles			<Article-processing charges – total if the Center is paying for these fees vs. authors incorporating into grant funding>
<Other>			

#### 4. CCAFS Data Management Strategy (2015-06)

- *Legal*
  - *Access Rights*
  - *Licensing of data and research*
  - *Open access/restrictions*
  - *Ownership of data*
- *Strategic Planning*
  - *Project level Data Management and Sharing Strategy/Plan*
- *Reference*
- *When*
  - *Decisions while designing*
- *PI*

Please find below the document

# CCAFS - Data Management Strategy

## Introduction

CCAFS is mandated to producing international public goods and has developed this Data Management Strategy (DMS) to enable the programme to fulfil its obligations with respect to making data and the relevant supporting documentation from its research activities available to the world community.

The Program Participant Agreements (PPA) established with CGIAR centres and other partners stipulate that data is to be made freely available and sets up the time scales for data publishing by scientists involved in CCAFS research activities:

“The Contracted Party agrees to publicly share any data and/or models generated as a result of activities under this Agreement through CCAFS’s data portals as soon as practically possible, but no later than twelve (12) months of generation for metadata and twenty-four (24) months for other data and/or models. Such data portals include, but are not limited to, the CCAFS agricultural trial data repository ([www.agtrials.org](http://www.agtrials.org)), the CCAFS climate data portal ([www.ccafs-climate.org](http://www.ccafs-climate.org)), the CCAFS Research Data on Dataverse (<https://dataverse.harvard.edu/dataverse/CCAFSbaseline>), and the repository of Agricultural Research Outputs (<https://cgspace.cgiar.org>). Access to the data should be fully granted to the CCAFS Knowledge and Data Sharing Unit at CIAT.”

The aim of the Data Management Strategy (DMS) is to guide the creation of an enabling environment where scientists and partners are able to produce and share high quality data outputs throughout CCAFS, while at the same time enabling a variety of data management procedures and practices at project level. This is achieved through creating “portals” specifically designed for common types of data where scientists can publish their data and by the provision of guidance and support to scientists and CGIAR Centres to facilitate producing well-managed and documented datasets that are easy to use both now and in the future.

Guiding principles for this strategy are:

- accessibility,
- ease of use,
- ethical use and sharing of personal and private data,
- provision of support for data generators,
- ensuring that credit and visibility go to data generators,
- adherence to international standards for data storage.

CCAFS aims to providing a “one-stop shop” for data generated by its research activities and expects to attract data contributions from scientists working in related areas even if not directly managed or funded by CCAFS. It will increase accessibility and visibility of scientific outputs to a global community for adding even more value to the products of CCAFS research with development outcomes in mind.

In this strategy we use the term “Data+” to indicate the actual data generated by the research process once it has been cleaned and is considered of good quality, as well as the documentation that will enable the use of these datasets in the future. This includes but is not restricted to documents about the methodology for data collection/generation, computer programs used for data manipulation and data processing, data quality assessment, and any metadata that helps in building a description of the context in which the data has been originated.

In defining this strategy, we have adopted the following principles:

*It has to be easy to implement and any burden to researchers that is generated from its implementation must be balanced by the benefits that the researcher will get from making his/her data available, and by the support that CCAFS will provide.*

*It should not affect the autonomy of scientists to carry out their research; the strategy ensures the independence and creativity of scientists in the collection of data that is relevant to the research objectives.*

## Goal

The goal of this DMS is for CCAFS data products to be archived and made available for long-term use by partners and the scientific community.

## Objectives

The objectives of this strategy are as follows:

1. To guide CCAFS is designing and implementing support mechanisms to reach the goal;
2. To make available quality-assured *Data+* to potential users now and well into the future;
3. To encourage appropriate levels of standardization, adoption of international standards and harmonization so that data from separate research activities can be brought together to enrich our understanding of processes, outcomes and impacts in the areas of the world where CCAFS works; and
4. To promote the production of “FAIR” outputs:
  - a. **Findable:** Data and metadata should be richly described to enable attribute-based search.
  - b. **Accessible:** Data and metadata should be retrievable in a variety of formats that are sensible to humans and machines using persistent identifiers.
  - c. **Interoperable:** The description of metadata elements should follow community guidelines that use an open, well-defined vocabulary.
  - d. **Reusable:** The description of essential, recommended, and optional metadata elements should be machine processable and verifiable, use should be easy and data should be citable to sustain data sharing and recognise the value of data.

## Scope

This DMS looks at making *Data+* available in public archives. It does not include research outputs such as papers and publications resulting from analysis of primary data. CCAFS has created alternative portals to share this type of information.

## Supporting mechanisms

Supporting mechanisms will be necessary for the implementation of this strategy. These include

1. Providing guidelines for making data available in such a way as to respect the trust that information providers have placed in CCAFS scientists;
2. Creating, maintaining and supporting portals that make data publication easy when CCAFS consider it necessary.

## Strategic Elements

### Programme Level

In order to achieve the objectives set out above, the CCAFS programme needs to:

- Based on Consortium level policies, discuss, define and adopt a data sharing and data ownership policy and Intellectual Property policy;
- Negotiate and coordinate actions with the Consortium Office of CGIAR (CO), as well as Participating Centres that are part of CCAFS;
- Include the required elements of these policies into the contracts established with Participating Centres – e.g. CCAFS Program Participant Agreements (PPAs);
- Put in place an implementation plan;
- Support and resource mechanisms to receive and archive data.

## Centre Level

In order to fulfil Participating Centre contractual obligations under the PPA agreements, CCAFS expects that centres will do the following:

- Allocate sufficient resources to allow for the implementation of the DMS;
- Utilise the provided support package for the implementation of the DMS;
- Ensure their data research outputs comply with the CGIAR Open Access Policy.

## Implementation

Three key elements are essential to the implementation of this strategy:

### 1. Establishing a *process*

A clear process for data sharing and management must be established, from legal agreements through to operating and reporting principles. CCAFS is implementing a planning and reporting system to enable the program to identify the data products that are to be generated and ensure that these products are made publicly available within the timeframes agreed with partners.

### 2. Supporting *compliance*

Support and encourage the use of data repositories that enable projects to comply with the CGIAR Open Access Policy.

### 3. Enabling a data *culture*

Implementing this strategy requires a significant cultural shift among program participants. Appropriate incentives and penalties should be established to promote data sharing. Metrics on data sharing from each program participant should be used as a criteria for measuring performance, reward or apply penalties. Among the conditions to facilitate the establishment of a more conducive data culture, CCAFS must:

- Support program partners in the process of submitting data to suitable repositories;
- Work with existing CCAFS repositories to enable interoperability;
- Highlight benefits to researchers to be derived from data sharing such as increased visibility, potential for increased collaboration and publication, and reputation;
- Make available statistics about data downloading and use so as to be able to use this information as a planning tool for the programme to promote CCAFS's research agenda and that of our scientific partners among the global audience.

# Data Ownership and Authorship

Main:

## 5. Data Ownership & Authorship

- *Communications*
  - *Data dissemination/publicity*
- *Strategic Planning*
  - *Data Ownership*
- *Legal*
  - *Ownership of Data*
  - *Partnership agreements*
- *Main*
- *When*
  - *Decisions while designing*
  - *Delivery of research products*
- *PI, Researcher*

Please find below the document



# DATA OWNERSHIP & AUTHORSHIP

## Introduction

This guide addresses questions such as why we need to have specific agreements about data ownership, who owns the project data, and who has the right to be named as an author.

## 'Data' in the context of Ownership

Throughout this guide we repeatedly use the term 'data'. By this we mean not only datasets, metadata, observational data, and statistical data, but also items such as reports, videos, images, maps and audio recordings.

## Who 'owns' the data?

Ensuring that each project has clear rules regarding data ownership makes sure that the relevant people have access to the data, whether these are scientists at the analysis stage or the public in the long term. These 'up-front' rules circumnavigate potential data ownership issues such as: researchers leaving projects and taking the only copy of the data with them, or a student refusing to share the data until after they have published etc.

## So, who 'owns' the project data:

- The project sponsors?
- The project leader?
- The institutions involved in the research?
- The individual researchers, the scientists, the fieldworkers?
- The project data manager or managers?
- The respondents who provided the data?
- The public?

In reality it is probably all of these.

At the outset of a project a 'data sharing agreement' or a 'memorandum of understanding' should be put in place. This document should clearly outline who 'owns' the data throughout the project process. This document should be shared amongst all project members and archived at the end of the project.

## Principles of Data Ownership

An example data sharing document is included in this Data Management Pack; it is based on the following 'ownership' principles:

- That research 'data' belong to institutions not individuals as it is only institutions that are able to ensure long term security for data;
- 'Data' generated by collaborating institutions belong jointly to those institutions;
- 'Data' collected using public funds are public property. Everyone has a responsibility to ensure the maximum value is realised from them.

## Rights and Responsibilities

A data ownership and sharing agreement should also cover the rights and responsibilities of the scientists with respect to project 'data', namely:

- Scientists generating research 'data' have a right to recognition for their work;
- Scientists generating research 'data' using public funds have a duty to use the 'data' for the purpose for which funding was provided and to publish the findings.

Intellectual property rights should be established in the contracts signed between the institutions.

In conclusion 'data' ownership and intellectual property should be managed in a way to balance the interests of individual scientists, their institutions, the donors and society as a whole.

## Authorship

There are many benefits to being a published author, in particular being recognised for your contribution to an area of research and being able to list your publications on your CV. However, being listed as an author does mean you are responsible for the accuracy of results, facts and interpretations within your publication, which you may be required to defend once peers have had the opportunity to review it.

### So, who has the right to be listed as an author?

1. Authors should make substantial contributions throughout the research process:
  - a. in the conception and design of the research, or
  - b. the analysis and interpretation of the data.
2. They should be involved in drafting the paper or critically reviewing it for intellectual content.
3. They should have approval of the final version prior to publication.

The table below summarises these conditions – to be an author you must fulfil each of these criteria.

<b>Authorship Criteria (must fulfil all 3)</b>		
<b>1</b>	<b>2</b>	<b>3</b>
Design OR Analysis OR Interpretation	Draft OR Critically Review the paper for intellectual content	Final approval prior to publication

Not including a project member as an author when he/she has been involved in each of these criteria denies them recognition for their work which they deserve due to their level of contribution.

### Who does not have the right to be listed as an author?

A project member does not have the right to be an author if they were not included in the design, analysis or interpretation stage, did not draft or review the paper, and did not have approval of the final version. For example, project members only involved in the data collection are not entitled to be authors, neither are members who manage the project without being involved in the actual activities.

Being included as an author if you have not fulfilled the criteria in the table above is dishonest, and may result in you having to defend work in which you have had little involvement, and which may be incorrect.

If a student carries out the work of part of a project and this is published as a thesis, then only the student is entitled to be the author of the thesis which they have written. Any publications resulting from a student thesis can be co-authored by project members meeting the authorship criteria; being a student supervisor does not automatically result in authorship rights.

## Editors

An editor must review, comment on, and approve the content of the whole paper. Editors are responsible for the content quality.

Simply making changes highlighted by others, or changing the presentation of a paper does constitute being an editor.

## Acknowledgements

Those who have contributed to the research but not in such a way as to be considered authors or editors are often listed under an “Acknowledgements” section. This is an established way of giving credit and thanks to those who have, for example, organized and managed the data, collected the data, devised the study tools, funded the research, run analyses, etc.

## Summary

The crucial points to take away from this guide are:

- A data ownership agreement for each project is essential – it protects the rights of everyone involved in the project from the respondents, through to the scientists, to the program funding the research; ensuring the data are used for the intended purposes and are available at the appropriate times to the appropriate people.
- Authoring papers has wonderful benefits, but also responsibilities. Ensure that you deserve the recognition associated with being an author and make sure that you are not excluding a member of the team who has contributed sufficiently to warrant authorship.

## Associate Videos

Videos accompanying the original release of the CCAFS Data Management Support Pack in 2013 are available as a playlist on the Statistical Services Centre YouTube Channel at <https://www.youtube.com/channel/UCs7EU95YMjhvNozJKCD92xQ/playlists>. These videos have not been updated since the original release but are mostly still relevant.

In particular the playlist includes a video on Data Ownership available from the following link: <https://www.youtube.com/watch?v=aDQWTuAMKTQ&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpi&index=5>

## 6. Data Ownership Agreement – Template

- *Strategic Planning*
  - *Data Governance*
  - *Data Ownership*
- *Legal*
  - *Ownership of data*
  - *Partnership agreements*
- *Using the Data*
  - *Sharing and Access internally*
- *Main*
- *When*
  - *Decisions while designing*
  - *Management of research processes*
- *PI, Researcher*

Please find below the document

# Data Ownership Agreement - Template

## Introduction

*[This template assumes that:*

- *A contract has been signed between partner institutions;*
- *Any intellectual property rights have been established in the contracts signed between institutions;*
- *Individual scientists and other data generators are subject to the terms of the contract signed by their institution for this project;*
- *The project is publicly funded or that a decision for making the data accessible to the public has been made by the leading institution.]*

This is an agreement between the signing institutions and scientists to establish data ownership and the right to use of data generated by the research activities of the research project *<project name>*.

It is understood that scientists and technicians working for the signing institutions are subjected to this agreement.

The agreement is based on the following principles:

- Research datasets belong to institutions not individuals as it is only institutions that can ensure long term security for data;
- Data generated by collaborating institutions belong jointly to those institutions;
- Data collected using public funds are public property; everyone has a responsibility to ensure the maximum value is realised from them;
- Scientists generating research data have a right to recognition for their work;
- Scientists generating research data using public funds have a duty to use the data for the purposes for which funding was provided and to publish the findings;
- Data ownership and intellectual property should be managed in a way to balance the interests of individual scientists, their institutions, the donors and society as a whole.

In practice these principles will be implemented as follows:

1. Institution *<institution name>* is responsible for co-ordination of all data management, sharing and accessibility activities. It will maintain databases of all data collected by the project and make the information accessible, according to the points below.
2. Each partner in this research project/programme agrees to supply all primary data and associated metadata collected as part of the project, to the co-ordinating institution for storage and archiving according to agreed activity milestones.

*[Activity milestones would vary according to the activity: for a short-term single activity it might be after data entry; for longer-term activities it might be after the data entry for each site.]*

The co-ordinating institution will provide instructions and support on the practical aspects of this process.

3. Primary data will be made publicly available by the coordinating institution <specify time> months after its generation.

*[For single data collection activities, such as a household survey, the data could be made available 12 months post generation. For long-term studies where data collection may be longitudinal, discussions should take place between the partners to decide whether data should be made publicly available at regular intervals throughout the collection process, i.e. every 12 months, or whether the data are not made public until the end of the data collection process, which may be several years. This is likely to be established by agreements established in the main contract between institutions.]*

4. Before the primary data are made public, project scientists will have a reasonable period of time in which to use the data for the agreed purposes.
5. The co-ordinating institution will safeguard the data and ensure that the scientists that have the right to access the data can do so both before and after the public release date.
6. Primary data will be made publicly available, ensuring that the generator/s of the data (the individual/s and associated institutions) is/are acknowledged as the author/s of the data. The data and the information about authorship of the data should be kept together.

## Definitions

### Agreed purposes

At the start of any project activity a protocol should be produced; this document includes details about the data that are to be collected and how they will be used (the agreed purposes), along with timelines for the activity which should include data collection timelines. All project partners should review and agree on the protocol prior to starting the activity.

### Data

'Data' in this template refers not only to: datasets, metadata, observational data, and statistical data, but also items such as: survey questionnaires, reports, videos, images, maps and audio recordings.

### Metadata

Metadata is information that fully describes the source and content of each dataset; it is information that allows a user to fully understand the dataset or resource. It also includes project level information that helps researchers to locate the data in a public archive. It is often described as "data about data". For further information please see the document "Introduction to Metadata".

### Primary data

By 'Primary data' we mean data that are of satisfactory quality to be of use. For numerical data this means the data should be clean and may include derived variables; for textual data such as reports, they should have been through the review process by the partners involved in the project. Primary data should also be anonymised as appropriate. It is the primary data rather than the raw data that will be archived. For further information see the document "Transition from Raw to Primary Data".

## 7. CGIAR Author Guidance

- *Communications*
  - *Data dissemination/publicity*
  - *External communications*
- *Legal*
  - *Licensing of data and research*
  - *Open access/restrictions*
  - *Partnership agreements*
- *Using the Data*
  - *Secondary users = others*
  - *Sharing and access to data*
- *Main*
- *When*
  - *Decisions while designing*
  - *Delivery of research products*
- *PI, Researcher*

Please find below the document

# Open Access: Publications

## Guidance for Authors

In November of 2013, all 15 members of the CGIAR Consortium unanimously adopted the CGIAR Open Access and Data Management Policy. The Policy commits CGIAR to making publications, data sets, and other final or completed versions of information products openly and freely accessible. CGIAR strongly believes that this clear commitment to Open Access will improve the efficiency, efficacy, and impact of its research and allow the global public to further benefit from it.

### Questions?

- Copyright transfers, negotiating with publishers, or potential legal obstacles to publishing (e.g. confidentiality or IP considerations): contact your local Center or CRP's Legal/IP Focal Point.
- Depositing manuscripts into the OA Repository: contact your local OA Focal Point.

### Three Steps to Comply:

#### 1 Retain your rights

Once manuscripts have been accepted by a journal, publishers typically require authors to sign a Copyright Transfer Agreement. CGIAR encourages all authors to use an Author Addendum to retain specific rights necessary to legally deposit and disseminate article manuscripts via repositories and comply with the CGIAR Open Access Policy. Of particular importance are the rights to deposit the peer-reviewed version of manuscripts, deposit immediately, and allow for re-use rights such as transitions.

#### 2 Deposit

As soon as the peer-review process has concluded and your article manuscript has been finalized, deposit a copy of the manuscript along with relevant metadata about the article into your Center's or CRP's repository. Include details such as the complete names of all authors, publication and journal information, abstracts, and appropriate keywords. CGIAR-specific details such as Centers and CRPs are also important. CGIAR Open Access Repositories are designed to make publications globally

#### 3 Share

searchable, discoverable, and accessible. Search engines such as Google and Google Scholar are able to index and provide access to materials in these repositories. While technology makes content available, using social media to announce articles' publication also helps spread the word, attract mainstream media attention, and make articles more widely discoverable.

### Copyright Transfer Agreements & Author Addenda

As part of the typical publishing process, publishers ask authors to sign a Copyright Transfer Agreement (CTA) or similar form. These forms ask authors to transfer copyright ownership – and all associated rights -- to the publisher.

Article 4.2.1 of the CGIAR Open Access & Data Management Policy requires that *"Peer-reviewed versions of scholarly articles...should be deposited in a suitable repository and made Open Access...When an author publishes in a closed access journal, he/she shall self-archive in an Open Access repository a digital version of the final accepted manuscript (the "post print" version)."*

**In order to comply with the CGIAR Open Access Policy, authors should deposit a final, peer-reviewed copy of each manuscript into a CGIAR repository and allow**

**for public access immediately** (ideally, otherwise no later than 6 months from the date of publication). It is critically important that you have retained the necessary rights from publisher to do so.

Increasingly, publishers are automatically granting certain privileges to authors such as permission to deposit. Even so, each journal and publisher has its own practices, many of which are inconsistent with the CGIAR Open Access Policy. Therefore, it is highly recommended that all authors attach a signed author addendum when signing copyright transfer agreements or any other publishing contracts in order to be compliant. CGIAR has prepared a model Author Addendum to use specifically for this purpose.

**Note:** The CGIAR Author Addendum aims to secure additional rights for authors

that exceed the minimum requirements indicated in the CGIAR Open Access Policy. For example, the Author Addendum aims to secure permission to immediately deposit the publisher's version of an article, not the peer-reviewed "post-print" version no later than 6 months from the date of publication. Authors are encouraged to consult their local IP focal point in order to consult the local IP focal point for assistance in negotiating with publishers and securing the appropriate rights.

### How to use the Author Addendum:

1. Review and sign a copy of the CGIAR Author Addendum.
2. Attach a signed copy to your publishing agreement.
3. Note in a cover letter to your publisher that you have included an author addendum to the agreement.
4. Mail the addendum to the publisher along with your copyright transfer agreement and cover letter.
5. If the publisher does not object to the Author Addendum, deposit the publisher's version of the manuscript into a CGIAR repository. If the publisher objects, consult your local Legal/IP Focal Point.

### Alternatives to using an Author Addendum:

1. Publish in a journal that automatically grants necessary permissions to authors. Review the journal's copyright transfer agreement or other relevant policies.
2. Negotiate with publishers to grant a license to publish, reproduce, and distribute the article, but ensure that authors retain all other rights, including the right to deposit the post-print, peer-reviewed article in an Open Access repository without a delay or no later than 6 months from the date of publication



## CGIAR Author's Addendum to the Publication Agreement

As a scientist of \_\_\_\_\_ (CGIAR Research Center), AUTHOR is bound by the CGIAR Open Access & Data Management Policy (<http://bit.ly/cgiar-oa-policy>) and relevant downstream institutional policies relating to the accessibility of publications accepted for publication.

**CGIAR** is a global partnership that unites organizations engaged in research for a food secure future. CGIAR research is dedicated to reducing rural poverty, increasing food security, improving human health and nutrition, and ensuring more sustainable management of natural resources. It is carried out by the 15 centers who are members of the CGIAR Consortium in close collaboration with hundreds of partner organizations, including national and regional research institutes, civil society organizations, academia, and the private sector. [www.cgiar.org](http://www.cgiar.org)

This addendum prepared in response to the CGIAR Open Access & Data Management Policy modifies and supplements the attached Publication Agreement ("Addendum") concerning:

---

(Manuscript title)

---

(Journal)

This Addendum and the Publication Agreement, taken together, allocate all rights under copyright with respect to all versions of the Article. The parties agree that wherever there is any conflict between this Addendum and the Publication Agreement, the provisions of this Addendum are paramount and the Publication Agreement shall be construed accordingly.

1. Author's Retention of Rights. Notwithstanding any terms in the Publication Agreement to the contrary, AUTHOR and PUBLISHER agree that in addition to any rights under copyright retained by the Author in the Publication Agreement, Author retains: (i) the rights to reproduce, to distribute, to publicly perform, and to publicly display the Article in any medium for non-commercial purposes; (ii) the right to prepare derivative works from the Article' and (iii) the right to authorize others to make any non-commercial use of the Article so long as Author receives credit as author and the journal in which the Article has been published is cited as the source of first publication of the Article. For example, Author may make and distribute copies in the course of teaching and research and may post the Article on institutional Web sites and in other open-access digital repositories.

2. Publisher's Additional Commitments. Publisher agrees to provide to Author within 14 days of first publication and at no charge an electronic copy of the published Article in a format, such as the Portable Document Format (.pdf), which preserves final page layout, formatting, and content. No technical restriction, such as security settings, will be imposed to prevent copying or printing of the document.

3. Acknowledgment of Prior License Grants. In addition, where applicable and without limiting the retention of rights above, Publisher acknowledges that Author's assignment of copyright or Author's grant of exclusive rights in the Publication. Agreement is subject to Author's prior grant of a non-exclusive copyright license to Author's employing institution and/or to a funding entity that financially supported the research reflected in the Article as part of an agreement between Author or Author's employing institution and such funding entity, such as an agency of the United States government.

For record keeping purposes, Author requests that Publisher sign a copy of this Addendum and return it to Author. However, if Publisher publishes the Article in the journal or in any other form without signing a copy of this Addendum, such publication manifests Publisher's assent to the terms of this Addendum.

AUTHOR	PUBLISHER
Name: _____ (corresponding author on behalf of all authors)	Name: _____
Signature: _____	Signature: _____
Date: _____	Date: _____

*The CGIAR Author's Addendum has been adapted from the SPARC Author Addendum (<http://bit.ly/sparc-addendum>).*

## Reference

### 8. Data Ownership Agreement - Example

- *Legal*
  - *Licensing of data and research*
  - *Open access/restrictions*
  - *Partnership agreements*
- *Reference*
- *When*
  - *Decisions while designing*
- *PI, Researcher*

Please find below the document

# Data Ownership Agreement - Example

## Introduction

This document is to define issues regarding data ownership and access of the research datasets to be generated during the lifetime of <name of project>, hereinafter referred to as “the Project”.

The Project is a collaborative venture between <name of collaborating institutes> and “the project team” is defined as comprising scientists and students based at one or more of the aforementioned institutes who are directly involved in the activities of the Project.

The Project is funded by <name of funding organisation>, hereinafter referred to as “the Funder”.

## Agreement

1. The data and protocols are intellectual property (IP) and hence are managed to conform with the requirements of the Funder.
2. Data and protocols are considered the joint property of the collaborating institutes.
3. Data may be used for two purposes:
  - a. Meeting the objectives specified in the protocols (called *Agreed Objectives*);
  - b. Meeting other objectives, including those not envisaged when the protocols were prepared (*Other Objectives*).
4. For each *Agreed Objective* in a protocol, there will be a statement of:
  - a. The individuals responsible for (or contributing to) the *Objective* (the *Agreed Participants*);
  - b. The output (report, paper, etc.) that will be produced;
  - c. The date by which the output will be ready (the *Ready Date*).
5. Until the *Ready Date*,
  - a. The *Agreed Participants* have exclusive rights over the data for the *Agreed Objectives*;
  - b. Other project partners may use the data for *Other Objectives* – in particular, the project data manager must have access to all the data for purposes of building and maintaining a project archive.
6. After the *Ready Date* all partners may use the data as they see fit, including making it available to third parties on the understanding that data have been anonymised as appropriate.

7. Authorship of outputs (articles, reports, etc.) will be determined as follows:
  - a. For outputs on *Agreed Objectives*, all *Agreed Participants* will be invited to be authors. Actual authorship will be according to contribution to the output.
  - b. For outputs on *Other Objectives*, authorship is determined by contribution to that output.
  - c. Student theses are solely authored by students.

## 9. CCAFS Publications Policy

- *Communications*
  - *Data dissemination/publicity*
  - *External communications*
  - *Support & guidance*
- *Reference*
- *When*
  - *Decisions while designing*
  - *Delivery of research products*
- *PI*

Please find below the document

# Publications policy

## CGIAR Research Program on Climate Change, Agriculture and food Security

The CCAFS publications policy covers

1. General principles
2. CCAFS publications series
3. Branding of CCAFS-funded research outputs

### 1. General principles

#### Open Access

- a) Peer-reviewed journal articles. CCAFS encourages publishing in open access journals<sup>1</sup>. **Gold open access** refers to the immediate availability of a publication free of charge on the publisher's or journal's website. Gold routes to open access include publishing in an open access journal (which is likely to charge an article processing fee) or through an 'author pays' (or 'hybrid') model which enables authors to publish articles in traditional subscription journals on an immediate open access basis following payment of a fee. This cost can be budgeted into CCAFS budgets. If researchers do not pursue gold open access, then they should pursue **green open access** where publishers permit you to submit a 'post-print' copy of an article to the CCAFS publications repository within six months of first publication<sup>2</sup>.
- b) CCAFS' own publications series, including Reports, Working Papers and Policy Briefs. All research outputs published under CCAFS series will have a Creative Commons license that encourages re-use with attribution<sup>3</sup>. Additionally, CCAFS-funded work published by CG centers and partners should be published using at least the same license, in order to facilitate easy co-publishing and wide accessibility.
- c) CCAFS CGSpace open-access repository. All CCAFS research outputs will be published into an open digital repository where they can be archived and re-used by others, in perpetuity. Researchers should send a) the final document suitable for publishing, in the case of workshop reports and other self-published outputs, including 'post-print' copies of peer-reviewed journal articles, as permitted by the publisher; or b) a web link to the output if it is already hosted in an open-access repository. This includes peer-reviewed journal articles that are published in open-access journals. Research outputs or URLs should be sent to [ccafs@cgiar.org](mailto:ccafs@cgiar.org) with the subject 'Publication submission'.

---

<sup>1</sup> This is in line with EU <http://bit.ly/9TQewP> and DFID policies <http://bit.ly/N4tgMw>

<sup>2</sup> You can verify publishers' copyright conditions as they relate to authors archiving their work on-line via the searchable RoMEO database [www.sherpa.ac.uk/romeo/](http://www.sherpa.ac.uk/romeo/). To comply with green open access as defined above, then the publisher must be green or blue in RoMEO,

<sup>3</sup> IRRI and ILRI have both adopted this standard. See more at <http://creativecommons.org/>

## Standards and styles

CCAFS has adopted the style guide for writers and editors developed by the World Agroforestry Center (ICRAF), available at the internal CCAFS planning site. This style guide includes the consistent forms of grammar, capitalization, punctuation, spelling, documentation, and language used in CCAFS publications. The standards should apply to all written outputs published under the CCAFS publications series and on the web (including the blog). Co-branded work published under another Center's series will conform to the Center's own standards as well as meeting those of the Center itself. The style guide may be downloaded from the CCAFS intranet: <http://intranet.ccafs.cgiar.org/SitePages/Publications.aspx>

## 2. CCAFS publications series

A Publications Committee has been set up consisting of the CCAFS Director and CCAFS Head of Research. Their role is to ensure that CCAFS produces a number of strategic reports and briefs in any one year, and that appropriate peer review (for different types of outputs) has been conducted and responded to by the authors. Full details of publishing processes are at <http://intranet.ccafs.cgiar.org/SitePages/Publications.aspx>.

The CCAFS publications series includes Policy Briefs, Reports, and Working Papers.

- Reports are longer and more detailed cross-cutting analyses.
- Policy Briefs are aimed at decision makers and development professionals, packaging research into concrete policy messages and recommendations.
- Working Papers are for works in progress, and are more technical.
- Workshop reports may be formatted using the CCAFS workshop report template.

The CCAFS Reports and Policy Briefs series will focus on cross-cutting syntheses. All publications are co-branded with the authors' research institutions, and will be jointly disseminated to key outlets, including online outlets, in person at events, and by direct mail, seeking input from the authors and partners to identify appropriate audiences and outlets.

## Types of publications in the CCAFS publications series

### (a) CCAFS report series

Containing important results and information for one or more of the stakeholder groups that we deal with. Peer reviewed. Funding from the central communications funds. About 5-10 reports per year maximum.

### (b) CCAFS policy brief series

Containing policy relevant messages for one or more stakeholder groups that we deal with. Peer reviewed. Funding from the central communications funds. About 5-10 briefs per year maximum.

### (c) CCAFS working paper series

Containing interim research results, or results from one part of an activity that will be eventually published elsewhere, and are not peer reviewed. Working papers must be approved by Theme Leaders or Regional Program Leaders. Funding from Theme or Regional budgets. CCAFS Coordinating Unit approves final product and assists with dissemination. **Note:** working papers may be published under a

partner institution's existing working paper series, following the guidelines for "Other CCAFS funded publications" described below.

Examples of outputs to be published under the working papers series include literature reviews, case studies, field-based research reports, and other intermittent research outputs.

#### (d) CCAFS workshop reports

A simple template has been developed for summaries and reports back from CCAFS workshops, to assist researchers who would like to produce these reports in a consistent format. Use is optional, and researchers may choose to use a workshop report template developed by one of the institutions participating in the CCAFS program.

### Process for Reports and Policy briefs in the CCAFS publications series

**Note:** A detailed estimate of production times for the CCAFS Report and Briefs is being developed and will be shared soon. This will help scientists plan and budget sufficient time for publication

- (a) Authors wishing to submit a report or brief should submit their proposal to the Communications manager ([v.meadu@cgiar.org](mailto:v.meadu@cgiar.org)) well in advance of the desired publishing date, so the Publishing Committee can make a decision as to whether to support the proposal. A proposal consists of the following information: Draft title, authors, draft key messages, key audiences for the product, key forthcoming events where hard copies should be available.
- (b) All Reports and Policy Briefs under CCAFS' own series are subject to an anonymous peer review. If the research has already been part of some other peer review process then the authors should demonstrate that.
- (c) After peer review, the document will be proofread, edited for style and language, and laid out in the CCAFS templates. This work will be coordinated and paid for by the CCAFS Coordinating Unit.
- (d) The Coordinating Unit will work with the authors and the communications staff at the authors' home institutions to identify communication and dissemination opportunities, and jointly disseminate the publication with partner centers.

### Process for Working Papers

**Note:** working papers may be published under a partner institution's existing working paper series, following the guidelines for "Other CCAFS funded publications" described below.

- (a) Authors must submit papers to Theme Leaders and Regional Facilitators for approval to publish as working paper.
- (b) Working papers are a numbered series. Authors must submit full paper title and list of authors to CCAFS Coordinating Unit (email [ccafs@cgiar.org](mailto:ccafs@cgiar.org)) to be assigned a working paper number. Working papers may also be submitted via the CCAFS intranet at <http://intranet.ccafs.cgiar.org/SitePages/Working%20Papers.aspx>.



- (c) The final document must be proofread and copy---edited as per CCAFS style guide, and laid out according to the CCAFS working paper template. The authors are responsible for overseeing this process and all costs will be invoiced to the theme or region (layout, proofreading, printing). The CCAFS Coordinating Unit can advise on consultants to assist with this if needed.
- (d) The authors should submit the final formatted working paper to the CCAFS Coordinating Unit for publishing in the open access repository. The Coordinating Unit will work with the authors to identify dissemination opportunities and jointly disseminate the publication with partner centers.

### ***3. Branding and acknowledgments***

CCAFS---funded research outputs that are not published using CCAFS---approved templates and are produced by the Centers/universities under their own---series, should be co---branded with the CCAFS identity and fully acknowledge CCAFS funding.

Such research products should be branded according to the CCAFS branding guidelines, which follows guidance set out by the CGIAR Consortium. Please consult the branding guidelines for full details of logo use and attribution:

<http://ccafs.cgiar.org/resources/branding---publishing---guides>

Program partners will make sure to inform the CCAFS Head of Program Coordination and Communications about significant upcoming publications in order to plan joint communications.

# Planning

## Main

### 10. Budgeting and Planning for Data Management

- *Managerial*
  - *Budgeting*
  - *Team composition/skill sets needed*
- *Strategic Planning*
  - *Project level Data Management and Sharing Strategy*
- *Main*
- *When*
  - *Decisions while designing*
- *PI*

Please find below the document

# Budgeting and Planning for Data Management

## Introduction

Data Management is a task generally acknowledged as necessary but not always budgeted for explicitly within a project. Frequently data management is done by each researcher in a project, to the best of his/her ability, independently of other researchers in the project. There are advantages and disadvantages to this approach. One of the main problems is to ensure data management adheres to the same principles and is of the same standard throughout the entire project so as to ensure the data are suitable for archiving and publishing. Even if this is achieved by multiple researchers within a project, the process is more likely to be efficient if carried out, supported or co-ordinated by a data manager.

As with any task within the project, data management requires an allocation of time, money and other resources. There is also the need to ensure the appropriate skills are available to the research process.

## The Data Management Role

The principal investigator needs to decide whether or not to appoint a data manager for the project. However, regardless of this decision, the functions of data management must be explicitly assigned to someone within the team. Appointing a data manager may be desirable when a project has multiple research processes going on simultaneously over a period of time, or when more than one researcher will require the support from someone who is technically competent with respect to managing the data.

The time and resources needed to effectively fulfil the data management tasks depend on the specific project size and complexity. In consequence, instead of giving an overall recommendation on the level of funding, we will establish the areas that we consider important to budget for.

## The Tasks of the Data Manager

The data manager supports and contributes to the research team's effort to gather, clean, make available, process, store and publish the team's research data and accompanying documentation. He or she has the responsibility to implement any agreements established by the research team with respect to the data and to ensure the team adheres to good data management practices. He or she is also in charge of establishing a system of data quality checks. In a separate document in this pack we describe the Terms of Reference for a Data Manager but have mentioned some of the tasks here as a reminder of the elements of Data Management that need to be resourced:

- Implementation of the data management plan;
- Set up and maintain a data and document storage facility (DDS);
- Perform quality assurance checks on the data;
- Provide support to the research team on data management;
- Set up data entry systems;
- Provide input into training;
- Provide input into the design of data collection tools;

- Prepare data and documentation for archiving;
- Collate the metadata;
- Archive the data.

## **Time and Budget for other Team Members**

Even if you have a full-time data manager, you will find that others in the team will also need to allocate some time to data management. In particular, the following areas should be included in your budgeting:

- Time for the principal investigator to oversee and co-ordinate the team data management efforts;
- Time for researchers to deal with data, data quality assurance, and data queries;
- Financial allocation for support staff or equipment for data entry;
- Financial allocation to establish a reliable system of backups. This might be to buy disk space on cloud servers, or it may be to buy and manage a more local backup system such as a local network drive. A combination of the two is not unreasonable.

## **Example list of Data Management Tasks**

Between September 2010 and December 2012, a Data Management Consultant logged 120 days of data management work for CCAFS. This included the following tasks:

- Formatting the questionnaire for the Household Baseline survey for ease of completion and data entry;
- Allocating variable names throughout the questionnaire;
- Providing support for users in CSPro including creating demonstration videos;
- Revising the CSPro data entry system to match major changes in the questionnaire after the pilot phase;
- Writing the manual for the data entry system;
- Writing a Data Checking Guide for use with the CSPro data detailing how to produce frequency tables within CSPro;
- Writing a guide for further data checks within SPSS and producing the corresponding SPSS syntax; the guide included instructions on how to transfer data from CSPro to SPSS;
- Drafting the Analysis Plan document;
- Producing the SPSS syntax to accompany the analysis plan;
- Documenting the procedure for running the analysis plan syntax and extracting the results;
- Running quality control checks on the data;
- Merging cleaned data files and recoding and consolidation of crop and livestock codes;
- Creating a Dataverse for the CCAFS Baseline Study;
- Uploading files to the Dataverse;
- Working on the mitigation questionnaires including setting up a spreadsheet for data entry and carrying out the data entry and checking;
- Checking site analysis reports including re-running analyses where necessary;
- Creating training videos on completing the household questionnaire and using the data entry system;
- Contributing to the CCAFS Data Management Strategy;

- Producing an assessment report on the quality of the data from the Household Baseline Study;
- Producing a version of the data entry system with screen labels in Spanish for use in Nicaragua and other Central American countries;
- Producing guidelines for data management;
- Etc.

The CCAFS Household Baseline Study was a “large” study – the questionnaire used covered 20 pages which resulted in a total of 970 variables. The survey was run in 15 core sites across 12 countries with 140 households from each site. Thus, the resulting data file had 2100 records. For a study of this size, employing a full-time data manager is not unreasonable.

## Summary

The debate about whether data management should be done by the researchers or by a dedicated data manager will no doubt continue. Some researchers have the capacity, the time and the inclination to do the data management themselves. However, in our experience, they are the exception. In most cases, support from a data manager who has been given explicit responsibility and authority to deal with data issues is essential to achieving the levels of quality that are expected from an international research effort.

## Associate Videos

Videos accompanying the original release of the CCAFS Data Management Support Pack in 2013 are available as a playlist on the Statistical Services Centre YouTube Channel at <https://www.youtube.com/channel/UCs7EU95YMjlvNozJKCD92xQ/playlists>. These videos have not been updated since the original release but are mostly still relevant.

In particular the playlist includes a video on Planning and Budgeting for Data Management available from the following link: <https://www.youtube.com/watch?v=O0vpXLJPB5o&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpj&index=4>

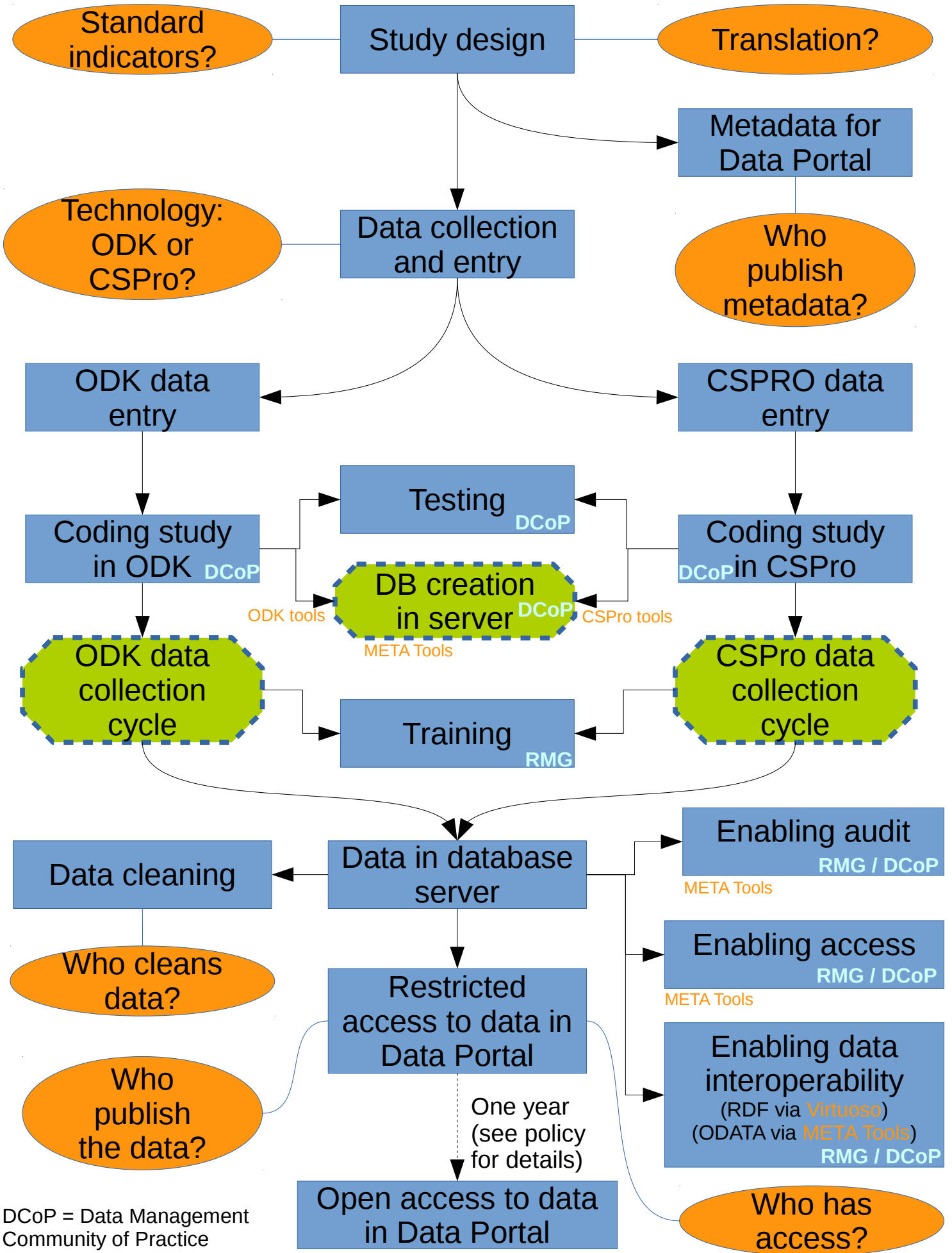
## 11. ILRI Data Management Process (flowcharts)

- *Technical*
  - *Choice of in-house vs external services/development*
  - *IT Systems (hardware, software, services)*
- *Data Management*
  - *Data Collection*
  - *Data Structures*
  - *Quality Control*
- *Managerial*
  - *DM System Definition/understanding*
  - *Informed selection of data collection tools/methods*
  - *Linking Data to Objectives*
- *Using the Data*
  - *Sharing and Access internally*
- *Main*
- *When*
  - *Decisions while designing*
  - *Management of research processes*
  - *Delivery of research products*
- *PI, Researcher, Technician*

Please find below the document

# ILRI's data management process

Author: Carlos Quiros

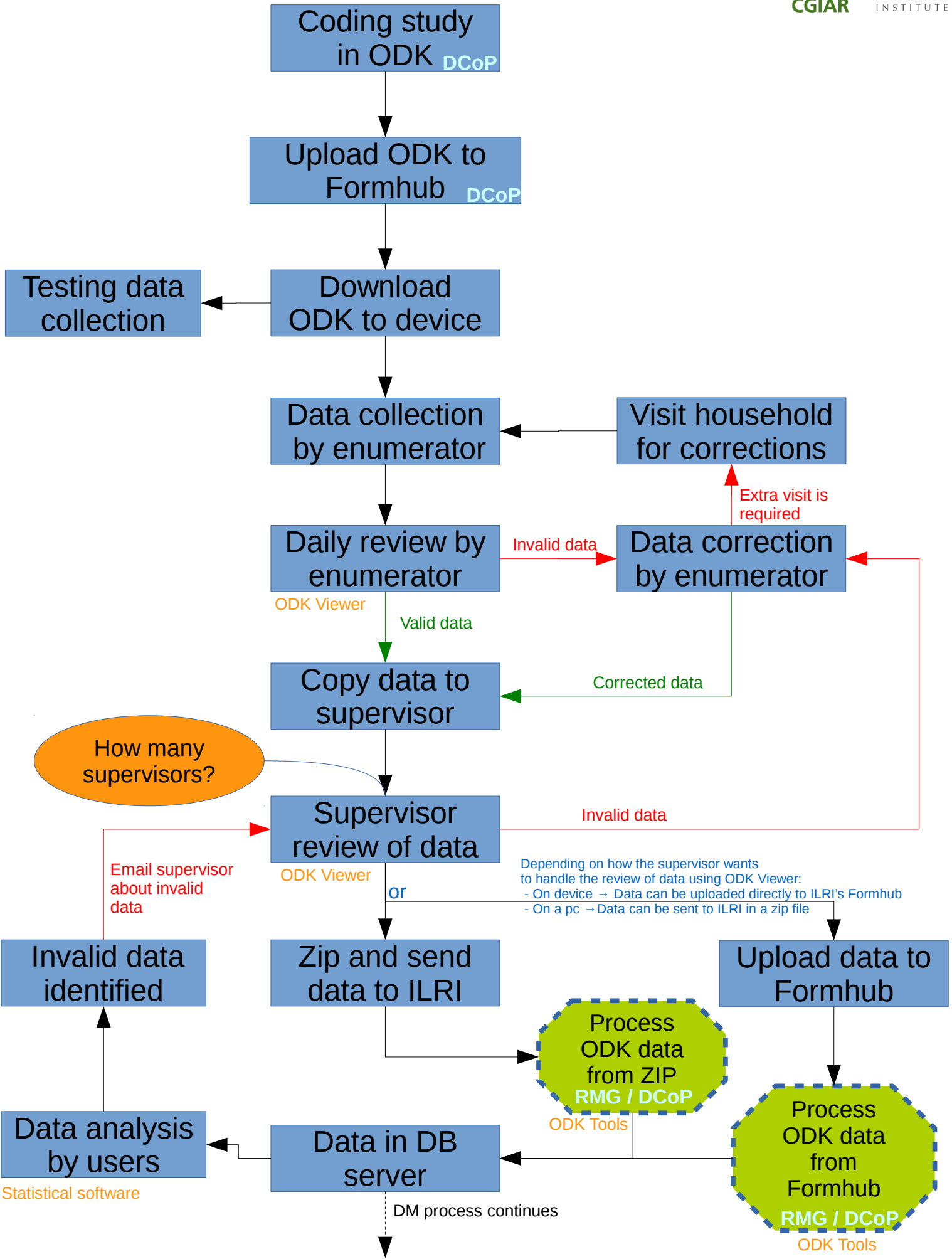


DCoP = Data Management Community of Practice



# ODK data collection cycle

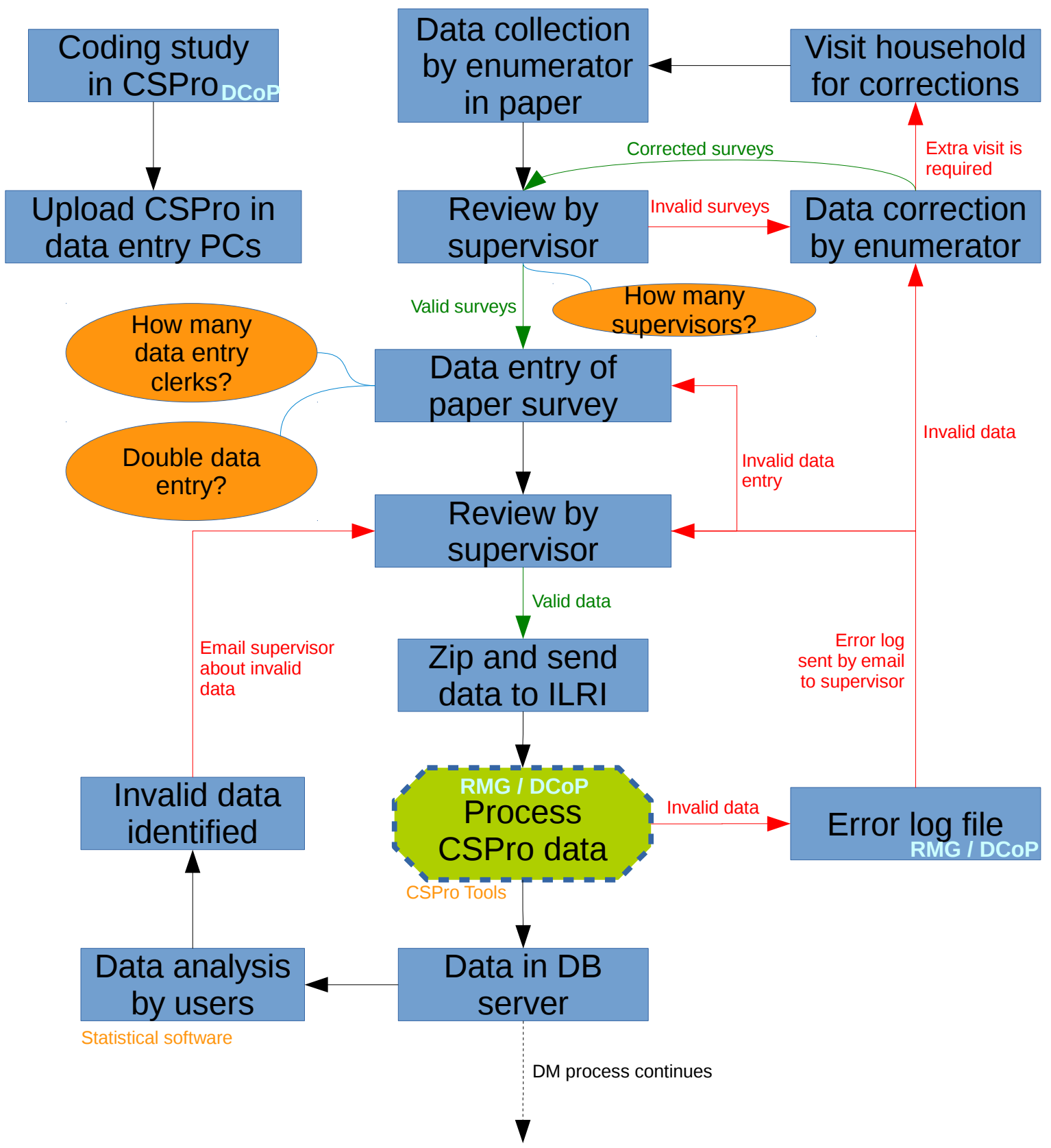
Author: Carlos Quiros





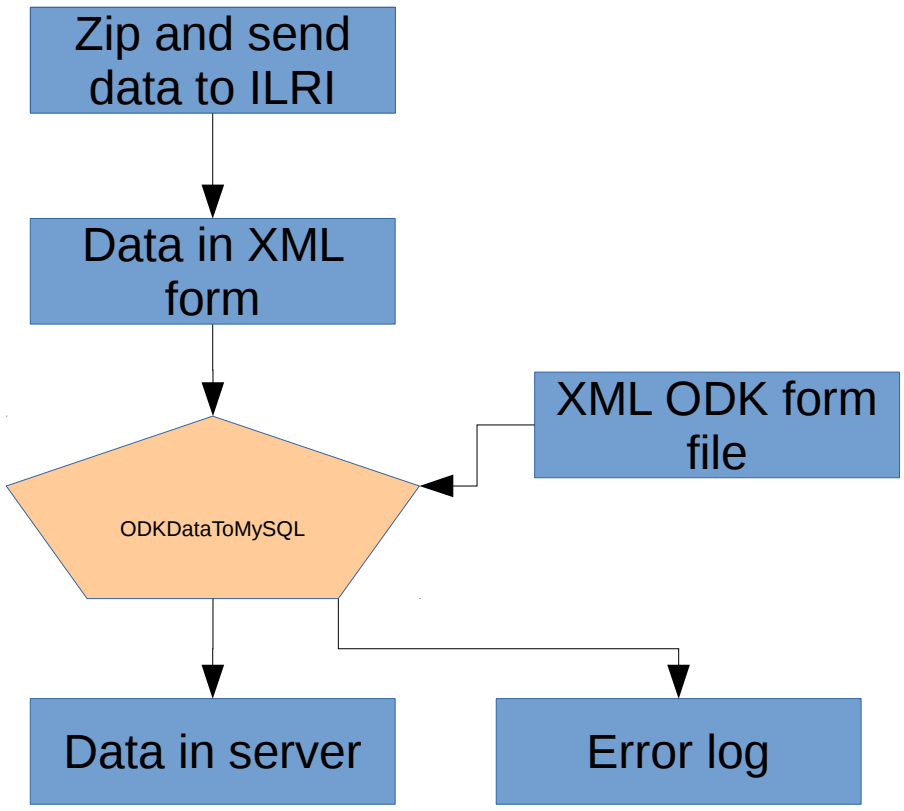
# CSPro data collection cycle

Author: Carlos Quiros



# Process ODK Data using Zip files (ODK Tools - internal to RMG or DCoP)

Author: Carlos Quiros

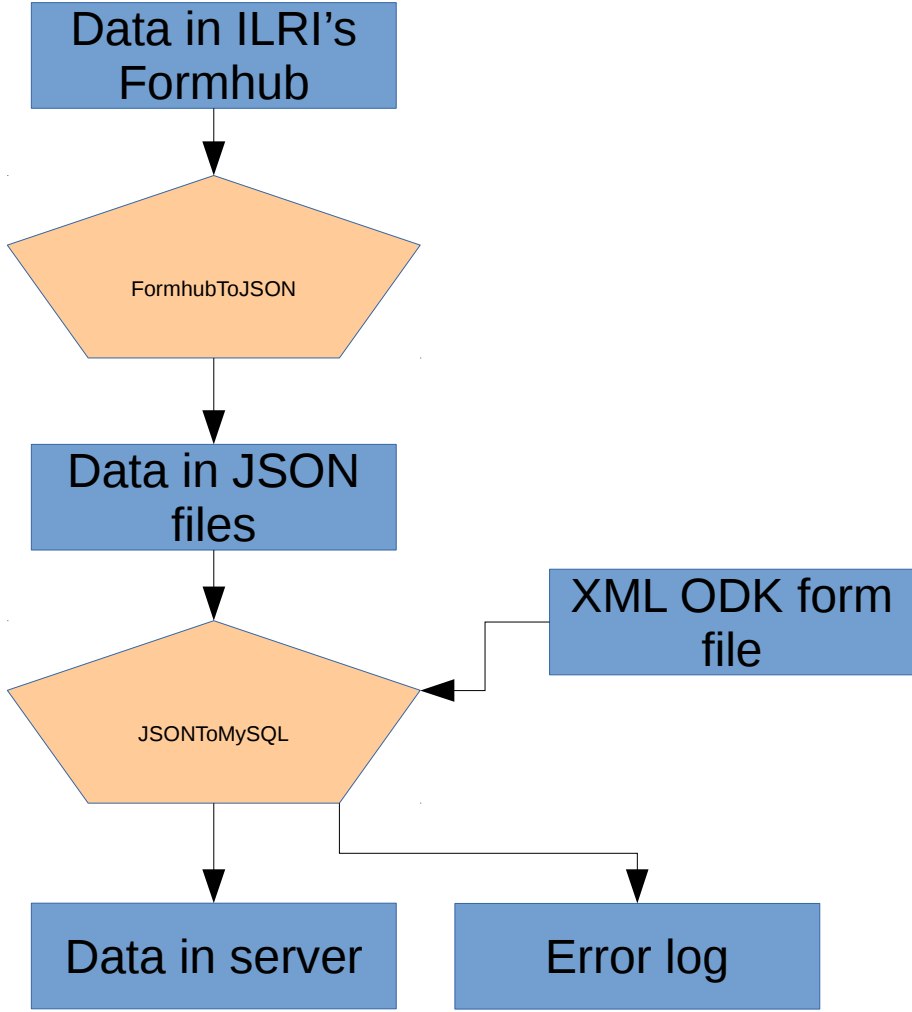


 = Software tool



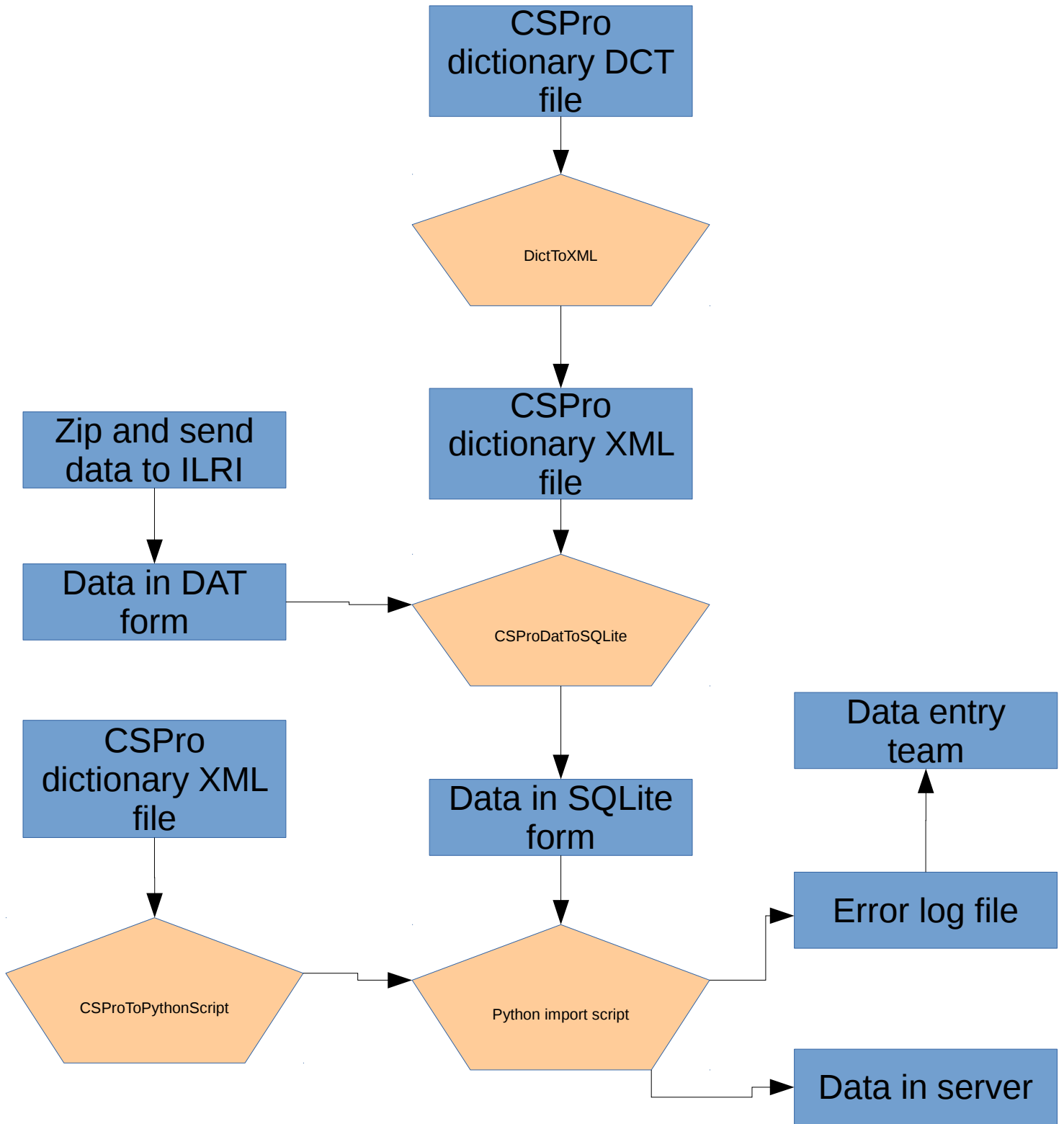
# Process ODK Data using FormHub (ODK Tools - internal to RMG or DCoP)

Author: Carlos Quiros



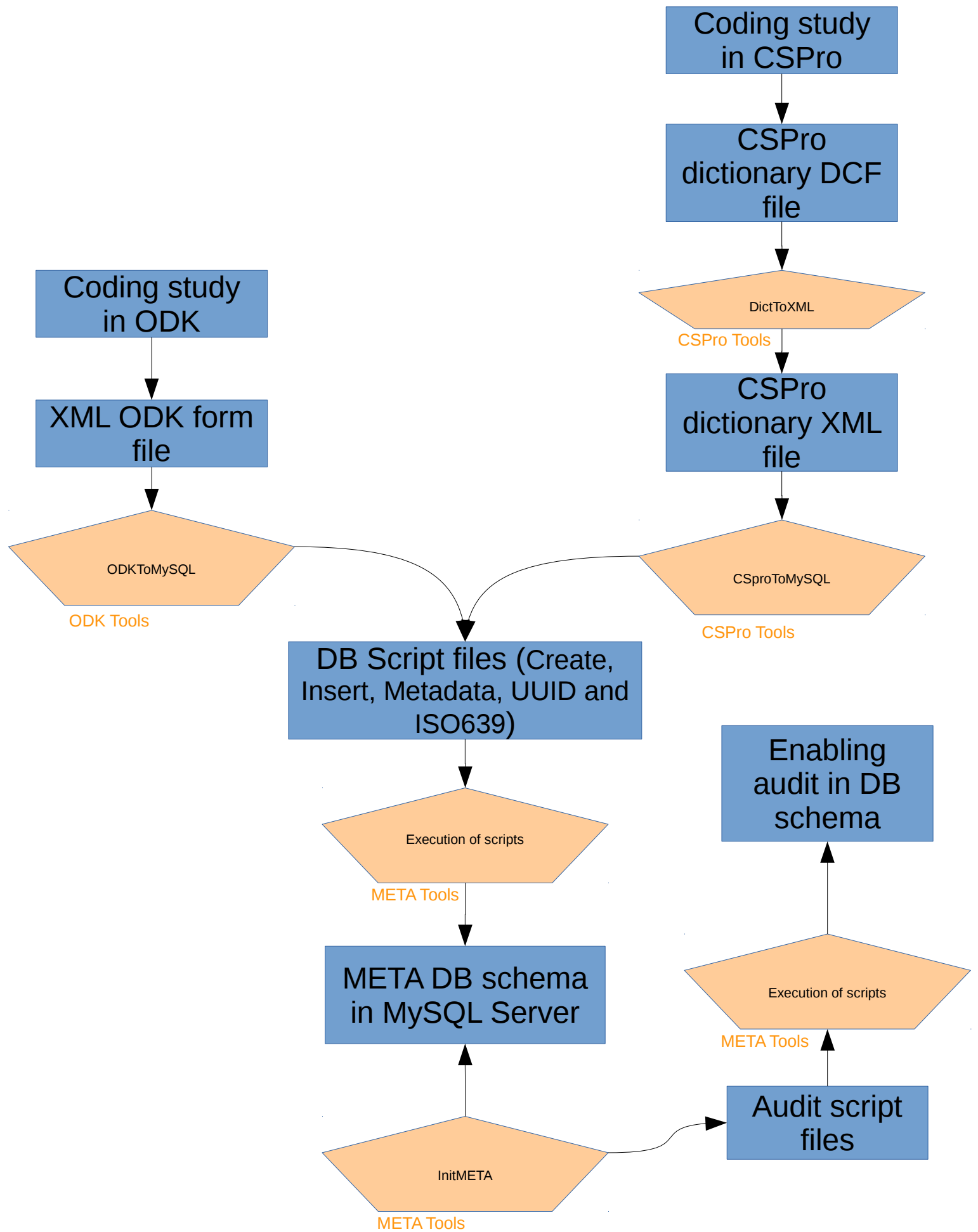
# Process CSPro Data (CSPro Tools - internal to RMG or DCoP)

Author: Carlos Quiros



# Database creation in server (internal to RMG / DCoP)

Author: Carlos Quiros



## 12. Data Management Process – Narrative

- *Technical*
  - *Choice of in-house vs external services/development*
  - *IT Systems (hardware, software, services)*
- *Data Management*
  - *Data Collection*
  - *Data Structures*
  - *Quality Control*
- *Managerial*
  - *DM System Definition/understanding*
  - *Informed selection of data collection tools/methods*
  - *Linking Data to Objectives*
- *Using the Data*
  - *Sharing and Access internally*
- *Main*
- *When*
  - *Decisions while designing*
  - *Management of research processes*
  - *Delivery of research products*
- *PI, Researcher, Technician*

Please find below the document

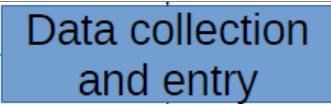
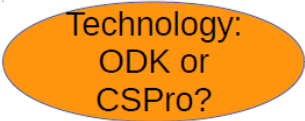
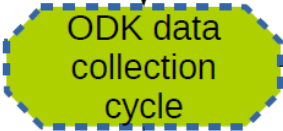
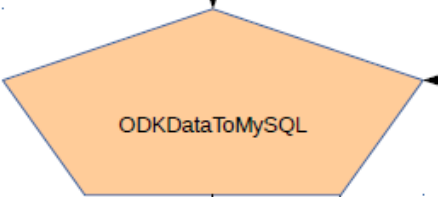
# Data Management Process Example Narrative

## Introduction

This document aims to provide some narrative to the flowcharts of the ILRI Data Management Process provided in a separate PDF document. The flowchart document was produced in 2014 and, although there have been some advances in data management practices since then, the charts are still very useful in reminding researchers of the various tasks involved in the whole data management process.

The flowcharts consider both digital data collection using ODK on hand-held devices, and data collection using paper questionnaires with the subsequent data entry using CSPro.

The elements of the flowcharts are colour-coded as follows:

Blue rectangles represent steps in the process	
Orange ovals contain questions that the PI and the project team should consider	
Green shapes with dotted blue edges link to later charts on the pages that follow	
Pale orange pentagons are software tools	

## ILRI's Data Management Process

Looking at the flowchart of the overall process we start at the top with the **Study Design** step and finish with archiving the data to a Data Portal in line with the CGIAR Open Access policy.

### Study Design

The study design stage includes setting the objectives of the study and working out what data needs to be collected in order to fulfil those objectives. In this process you will be developing the questionnaire to be used. Questions at this stage include deciding on what standard indicators are going to be used and whether or not translations are needed. Where possible it is ideal to have the questionnaire in the local languages and to do back-translations to ensure the original meaning of the questions is not lost in translation. This is far preferable to on the spot translations by the enumerators during the interviews

## Metadata

Now is the ideal time to start pulling together the metadata that will be needed when the data are archived; if you leave this task until the end then those with the necessary information are likely to have moved on to other projects and gathering the necessary information becomes very difficult and time-consuming. Think about where the data are to be published and pull together the necessary information. Remember this is the project level metadata which could be referred to as the Study Catalogue; later on, there will be the description of the dataset itself which will also need to be included – we refer to this as the Data Dictionary. See the guide “Introduction to Metadata” for further information about the levels of metadata.

## Data Collection and Entry

We then move onto the process of collecting and entering the data. The key question at this stage is about the technology to use: are you going to use digital data collection using ODK or are you going to use paper questionnaires and enter the data using CSPro? Of course, it is entirely feasible that you use digital data collection in some sites and paper questionnaires in others.

## Coding the Study

Once you have decided which technology to use the next task is to create the ODK system or the CSPro system (or both). In theory data collection using paper questionnaires could be started before the CSPro system is prepared; however, it is a good idea to have the system working before data collection starts so that data entry can start as soon as some completed questionnaires become available. Of course, it is essential to have the ODK system ready before fieldwork can start.

## Testing

Both the ODK system and the CSPro system must be thoroughly tested. Ideally you will be carrying out a pilot study and a pilot is the perfect opportunity for road-testing the systems. Remember both systems should include automatic checks on the data and we would recommend compiling a comprehensive list of checks to incorporate into the system. These checks might include checking that particular values are within a specified range or checking for consistency across variables. For example, in a household roster you would want to avoid situations where a child appears to be older than his/her parents. Some checks might be “hard” checks while others might be “soft” checks. For example, you might expect a mother to be at least 15 years older than her child but accept that some mothers might be as young as 10 years of age; you could therefore have a “hard” check whereby the system will not accept a mother being only 9 years older than her child and a “soft” check where you ask for confirmation if the age difference is between 10 and 15 years.

## DB Creation on Server

Once you know all the variables you will be creating you can create the database on the server to hold the collation data. We will expand on the database creation process later.

## Training

In the flowchart, training appears to follow the data collection. However, training should come before data collection. Regardless of whether you are using ODK or CSPro the enumerators will need to be trained to carry out the fieldwork. They will need to become familiar with the questionnaire and the types of responses that are expected – for example which questions only accept a single response and which are multiple response, which require a date, and which are open text questions. If you are using ODK, the enumerators will also need to be trained on using the ODK system. Separate training for the supervisors would cover tasks such as uploading data to the aggregate server, etc. If you are only



using paper questionnaires, the enumerators will not need to know how to use the CSPro system. Instead, you will need to train data entry staff in the use of this system.

## **ODK & CSPro Data Collection Cycle**

After the training we have the data collection phase. Data collection in ODK and in CSPro will be expanded on in the next main section.

### **Data in Database Server**

Towards the end of the data collection cycle, the data will be collated into the database created earlier on the server.

### **Data Cleaning**

There is then the process of data cleaning and the main question here is “Who cleans the data?”. It must be remembered though, that data cleaning is not just a single phase; errors can appear in the data at various points in the project life-cycle, so it is important to remain vigilant. We would also recommend keeping an audit trail of changes and corrections made to the data. The audit trail should be included with the data archive so that you can defend your results should anyone question the validity of your data.

### **Enabling Audit, Access & Data Interoperability**

These are all tasks related to the Research Methods Group (RMG) and Data Management Community of Practice (DCoP) at ILRI and is about allowing these groups access to the database so that they can audit the data and integrate it into their systems.

### **Restricted Access to Data in Data Portal**

Set out in the policy documents for the organisation, there will be mention of the length of time researchers on a project can have restricted access to the data for them to be able to work on publishing their results. This is standard practice and is generally something like 12 months from data collection. At this stage you will need to determine who should have access and who is going to publish the data.

## **ODK Data Collection Cycle**

### **Upload ODK to Aggregate Server**

Once the study has been coded in ODK the next step is to upload it to the aggregate server. In the flowchart “Formhub” is mentioned; Formhub was an aggregate server that was popular a few years ago but there are now many others such as ONA.

### **Download ODK to device**

Once the ODK system is on the aggregate server you need to download it to the hand-held device ready for the fieldwork. Of course, you must already have installed ODK Collect onto the device and will need to have adjusted the settings to link to the aggregate server that you are using. We would suggest that only the supervisors download the system to the devices to ensure all enumerators are using the correct version.

### **Testing Data Collection**

Once the system has been downloaded it will need to be checked on the device. If problems are found or changes are needed then you will need to go back to the coding stage, upload the revised form and

download it again to the device. It is important that all enumerators have the same working version of the ODK forms on their devices.

### **Data Collection by Enumerator**

This is where the enumerators interview household members and collect the data. The number of households that can be visited in one day will depend on the length of the questionnaire and this will have been determined during the training.

### **Daily Review by Enumerator**

At the end of each day there should be time for the enumerators to review the data they have collected and check for any invalid or inconsistent data. Many of these are likely to have been found by the ODK system itself but it is not always possible to check for every eventuality, so a review is still necessary.

### **Data Correction by Enumerator**

If invalid data is found, then the enumerator will need to correct this.

### **Visit Household for Corrections**

It may be the case that the enumerator can correct the data without revisiting the household. For example, he/she may notice a spelling mistake in the name of the village. However, there may be errors or inconsistencies that require a revisit to the household to confirm or correct values.

### **Copy Data to Supervisor**

Whether or not corrections were needed after the enumerator review, only valid or corrected data should be passed to the supervisor. You will need to decide how many supervisors you need but of course this should be decided before the start of the fieldwork.

### **Supervisor Review of Data**

Once the supervisor has the data from the enumerators, he/she should do his/her own review. If invalid data values are identified, then these should be returned to the relevant enumerator for correction. It may seem excessive for both the enumerator and the supervisor to review the data, but as we said earlier on, the process of data cleaning and checking is an on-going process and errors can be found at any stage.

### **Upload Data to Aggregate Server**

Once the supervisor is satisfied that the data are valid, he/she can then upload the data to the aggregate server. We suggest that only supervisors upload data as this helps to ensure the validity of the data.

### **Zip and Send Data to ILRI**

An alternative to uploading the data to the aggregate server is to zip the data and send it directly to ILRI to be processed. This is also after the supervisor review and any subsequent corrections.

### **Process ODK data from Aggregate Server or ZIP file**

In a separate section we look at the stage of processing the data either from the ZIP file or from the aggregate server.

## Data in Database on Server

Once the data have been processed it will be collated into the database on the server that was created earlier.

## Data Analysis by Users

At this stage the researchers can start on the analysis of the data. During this process they may identify invalid data.

## Invalid Data Identified

If invalid data values are identified during the analysis, then data queries should be sent back to the supervisor who will need to review the problems and arrange correction if possible. Of course, by this stage it might not be possible to revisit households to check on data, so a decision will need to be made about any errors and inconsistencies. It is important to document these decisions.

## CSPRO Data Collection Cycle

For data collection using paper questionnaires and subsequent data entry in CSpRo, the process of coding the study in CSpRo (i.e. preparing the data entry system) can be done in parallel with the fieldwork. However, we would suggest you aim to have the system completed and tested before any completed questionnaires start coming in so that data entry can be started straight away.

## Coding the Study in CSpRo

This step includes thoroughly testing the system. As mentioned earlier, many data checks and question skips can be automatically programmed into the system, and all such skips and checks need to be tested.

## Upload CSpRo onto Data Entry PCs

The final CSpRo system will need to be placed on all PCs being used for data entry. You would need the CSpRo software installed on the PCs and, from the data entry system itself, you would need the dictionary file (extension .dcf) and the compiled binary version of the system (extension .pen or .enc for earlier versions of CSpRo).

## Data Collection on Paper

Meanwhile data collection can be taking place. It is essential that the enumerators have been trained to carry out the interviews and know what is expected of them in terms of completing the questionnaires. Everything entered on the questionnaire must be clear; any errors should be crossed through neatly, so it is obvious that this is a mistake. We recommend having a section on the front of the questionnaire with space for enumerators, supervisors, and data entry staff to sign and enter the date. In signing the questionnaire, they are taking responsibility for having completed their tasks.

## Review by Supervisor (paper questionnaires)

Before the completed questionnaires are sent for data entry, the supervisor should visually review the data looking out for any inconsistencies or other errors. It is useful to have a checklist for this process so that nothing is missed. You will need to decide on the number of supervisors you have. This is likely to depend on the size of the team. Note that the data entry system will catch some inconsistencies, but it is important to check the completed questionnaires while still in the field so that there is the possibility of returning to the household to double-check the data.

## **Data Correction by Enumerator**

Where errors are found the corresponding questionnaires are returned to the relevant enumerators for correction and checking. Once corrected the questionnaires will again go to the supervisor for review.

## **Visit Household for Corrections**

For some errors or possible inconsistencies, a return to the household is desirable. Corrections should be clearly marked on the questionnaires and once again passed to the supervisor for review.

## **Data Entry of Paper Survey**

Data entry should be done as soon as possible after data collection. There is then a greater chance of being able to correct any errors that might be found during data entry.

When you are thinking about the data entry you will need to decide on the number of data entry clerks you are likely to need. Remember the data entry staff will need to be trained in the use of the CSPro data entry system. You also need to decide whether or not you will be using Double Data Entry (DDE). This is where two different data entry clerks enter the same data into two different files. The files are then compared, and any differences are checked against the paper questionnaires. CSPro includes a Data Compare tool to facilitate the data comparison.

## **Review by Supervisor (data in CSPro)**

Once the data have been entered they should be reviewed again by the supervisor and corrections made where possible.

## **Zip and Send Data to ILRI**

Valid data are then zipped and sent to ILRI for processing.

## **Process CSPro Data**

Tools within CSPro are then used to process the data. This process is expanded in a separate section.

## **Error Log File**

If errors are found in the data, then a log file is produced which is emailed back to the supervisor for review or back to the enumerator for correction.

## **Data in Database on Server**

The step of processing the CSPro data leads to the data being stored in the database that was created earlier on the server.

## **Data Analysis by Users**

At this stage the data should be ready for the researchers to start some analysis using whatever statistical software they feel is appropriate.

## **Invalid Data Identified**

During the analysis it is possible that further errors or inconsistencies in the data will come to light. If this is the case, then the supervisor should be contacted, and corrections made. If it is not possible to correct or check the data at this stage, then a decision needs to be made about whether to include these particular data values in the analysis or to treat them as missing values. Whatever decision is made it should be documented along with the reasons for that decision.

The remaining flowcharts in the accompanying file give details about the processes at ILRI. They are useful here as examples of how the data might be processed and below we briefly summarise each of these steps.

### **Process ODK Data using Zip Files**

When the data are sent to ILRI in Zip Files it is expected to arrive in XML format. The ODK form itself is also in XML format and both the data and the form are fed into a software tool which converts ODK data to MySQL. The two outputs from the software tool are (i) the data in the database on the server, and (ii) an error log. The error log, as reported earlier, would go back to the supervisor for review and data corrections.

### **Process ODK Data using Formhub**

The process here is very similar to the previous process. In this process there are two software tools: the first converts to the data from Formhub into JSON which the second converts JSON to MySQL. The second software tool also takes as input the ODK form in XML format. Here again the result is data in the database on the server and an error log.

### **Process CSPro Data**

For the CSPro data, the dictionary (stored in the .dcf file) is converted to XML format. Meanwhile the data is taken from the zip file as an ASCII file (with the extension .dat) and, together with the XML version of the dictionary, is fed into a software tool that converts CSPro data to SQLite. The XML version of the dictionary file is converted to Python Script, and this, together with the SQLite version of the data, goes through a Python import script which stores the data in the database on the server and produces an error log which goes back to the data entry team.

### **Database Creation on Server**

The database itself is created by first converting the ODK form and the CSPro dictionary to MySQL. Various scripts are run to add in the metadata and unique identifiers and basically prepare the database to be populated when the data have been collected and checked.

### **Summary**

As we have mentioned, the database creation and the processing of the data to transfer it to the database are particular to ILRI but the flowcharts give an idea of the sort of processes that are required.

The overall data management process and the data collection cycles in both ODK and CSPro are nicely detailed and in this narrative, we have tried to include additional comments and things that should be considered at each stage. The processes include a lot of data checking and reviewing, and some might be tempted to think that this is excessive. However, as we have mentioned, errors can creep into the data at any stage, so it is important to remain vigilant so that you have a dataset that is as free from errors as possible for your analysis.

### 13. Creating a Data Management Plan

- *Strategic Planning*
  - *Data Ownership*
  - *Project level Data Management and Sharing Strategy/Plan*
- *Main*
- *When*
  - *Decisions while designing*
  - *Management of research processes*
- *PI, Researcher, Technician*

Please find below the document

# Creating a Data Management Plan

## Introduction

The main purpose of a data management plan is to help you produce high quality data thus reducing the risk of producing results that are not robust. A general aim of research is to improve the quality of individuals' lives; for example, by reducing poverty, promoting development, relieving pain, etc. Unreliable results are not useful for this purpose. By considering and detailing the data management tasks needed throughout the research project, you can ensure you have the necessary resources in place in terms of time, people skills, equipment, and finances. A data management checklist will help ensure that nothing is overlooked.

This document is aimed initially at the Principal Investigator (PI) whose task it is to allocate data management responsibilities to a member of the team. The person with data management responsibilities will be referred to as the "Data Manager".

## Project Plan & Activity Protocol

In this document we will consider two levels – the project level and the activity level. At the project level the "Data Management Policy" aims to establish principles and agreements concerning data generated by the project as well as considering the resources that will be needed. At the activity level a "Data Management Plan" details the specific procedures that will be carried out to put the policy into operation. There should be a data management plan for each research activity. The PI has responsibility for drawing up and ensuring implementation of the overall policy while the Data Manager is responsible for activity level plans and procedures.

For example, it is CCAFS policy to archive data generated from project activities. This could be expressed in the CCAFS Data Management Policy as:

*Data generated by CCAFS will be submitted to a public archive within 24 months of collection.*

Each activity level plan should then detail the steps that will be taken to prepare the data for archiving including anonymisation, producing the data dictionary, etc. The plan will also say who is responsible for each task.

## Project Level Plan

The Project Data Management Plan would generally be drawn up by the PI. By the time of writing the plan, the PI should have a good idea of what activities are to be carried out and should therefore be able to make decisions about resources and the allocation of data management responsibilities.

## Do I need a Data Manager?

One of the first decisions to be made is whether a data manager is needed for the project. The choice is basically between:

1. Having a specialist data manager to whom all data management responsibilities are allocated;
- or

## 2. Allocating data management responsibilities to scientists.

The decision depends on the size and complexity of the study and the skills of the scientists.

*Most projects that involve a team of scientists (as opposed to a single researcher) are likely to need a data manager.*

We would recommend the inclusion of a data manager in most projects. This may correspond to the allocation of data management responsibilities to an existing member of the team who has the relevant skills, time and inclination to do the job well, or may involve recruiting a new member of staff with the relevant skills. A separate guide exists listing the terms of reference for a project data manager.

### Principles & Agreements

As already mentioned the data management policy aims to establish principles and agreements with respect to data generated by the project. These should include:

- Data ownership
- Data sharing & access
- Ethics, privacy and copyright
- Archiving
- Quality Standards & Security
- Resources & Responsibilities

### Data Ownership

Data ownership is often a contentious issue and it is important to draw up agreements from the outset to avoid problems later. All researchers will need to sign up to these agreements. Data ownership is covered in more detail in a separate guide. As previously mentioned, CCAFS is now governed by the CGIAR Open Access Policy and the principles of this policy should be made clear to all researchers working on CCAFS projects so that they understand from the outset that they cannot keep the data to themselves.

### Data Sharing & Access

This is linked to data ownership in that if it is established that all data are owned by the project, it follows that all team members must have access to the data. However, there are issues regarding confidentiality, so it might be that only a limited few have access to the raw data, but others can access the anonymised data, or that data access occurs in a staged manner with more individuals being given access with time. Thus, the decisions to be made are:

- Who has access?
- Who grants access?
- How is data accessed?

### Ethics, Privacy & Copyright

If the research involves collecting data on individuals the researcher would generally need to obtain ethical approval or establish the code by which the project will work; respondents must be fully informed about the purpose of the study; their data cannot be included unless they have given informed consent and they must be given the option of withdrawing from the project at any time; all personal data must remain confidential – in general individuals should not be identifiable from any



data that is put into the public domain although there may be exceptions; GIS references pose a challenge to this. Copies of the consent forms should be included in the project archive.

If any copyrighted data or data collection tools are to be used, then permission must be sought from the copyright holder. Any legal restrictions should be considered.

### **Archiving**

This is likely to be a broad statement in the data management policy stating for example that all data generated by the project will be put into the public domain within a specified timeframe. It would then be the responsibility of the data manager to draw up a document describing the requirements for archiving which scientists must incorporate into the data management plans for their activities.

### **Quality Assurance & Security**

Where quality assurance and security are concerned the principles stated in the policy document should prompt the data manager to include details under these headers in the activity level plans. For example, the project policy document might include the statement:

*Project data will be subject to a quality assurance process*

The activity data management plans should then detail the steps to be taken. For example:

- CPro will be used for data entry and a customised data entry system will be produced;
- Automatic skips will be programmed into the system which follow the skips in the questionnaire;
- Double-data entry will be used and <the data manager> will carry out the data entry comparisons and subsequent corrections;
- Backups to external hard drives will be taken regularly with incremental backups taken at the end of each day and full backups once a week. <The data manager> is responsible for ensuring that these backups are done;
- Etc.

### **Resources & Responsibilities**

Resources include people (skills & responsibilities), equipment (hardware & software), time and money. The PI should be aware of the resources currently available and, by examining the list of planned research activities, should be able to draw up a list of requirements adding in some element for contingency. The plan should clearly state who is responsible for the data management of the project.

In the box below, we have set out a draft template for a project data management plan.

### Data Management Policy for <Project Name>

This plan sets out the data management principles and responsibilities for the <Project Name> research project. For each research activity a data management plan must be produced. These plans should detail the steps that will be followed to ensure these principles are adhered to.

**Data Manager:** The data manager for the project will be <name of DM>. He/she will have overall responsibility for ensuring the timely completion of all data management tasks.

**Data Ownership:** All data generated by the project will be owned jointly by the collaborating institutions <names of collaborators>. A data ownership agreement will be drawn up and signed by all parties.

**Data Sharing & Access:** All scientists in the project team must have access to the data as and when needed.

**Ethics, Privacy & Copyright:** Ethical approval must be sought for research activities that involve human subjects. Confidentiality of personal data must be maintained. Any copyrighted materials used in the research must be acknowledged correctly.

**Quality Standards & Security:** All project data will be subject to a quality assurance process. Regular backups will be taken throughout the project.

**Archiving:** All data generated by the project will be archived within 24 months of data collection

## Activity Level Protocol

The activity level data management plan would be drawn up by the data manager or the person with data management responsibilities within the team. This is a much more detailed document explaining how the principles are going to be achieved. For example, how they intend to set up a data and document storage facility for data sharing among the team.

The plan will naturally vary according to the type of activity but would generally include the following elements:

### Data Collection

#### Capture Methods

Briefly describe the activity.

- Will you be running survey(s) or conducting experiment(s)/trial(s) to collect your data? Describe the methods to be used.
- What technology will be used for capturing the data – paper, mobile device, etc.?

#### Data Description

Include a brief description of the information to be gathered including the nature, scope and scale of the data that will be generated. For example:

- What are the study units and how many will there be? For example: we will be collecting data on 20 households in each of seven villages.
- What mechanisms do you have to check that the correct amount of data is collected – i.e. that you have the right number of study units?
- Where appropriate have you determined the units of measurement to be used? How will you ensure consistency in the units used?

## Secondary Data

Is there a need to review secondary data in your project activity? If yes, then describe the data to be used.

- Where are these data currently stored?
- How are they to be accessed?
- Who owns the IPR on these data?
- Can new data generated by the current activity be easily linked to the secondary data – what are the linking fields?

## Computerisation & Storage

### Data Entry

- How will you capture the data – on paper or directly onto hand-held devices or laptops?
- What software will you be using for data entry?
- Have you a customised data entry system or will you be preparing one? What data checks are/will be included in the system?
- Will the data entry system be documented?
- If recording directly onto hand-held devices what mechanisms do you have to ensure quality? For example: two researchers will be present during data collection to validate values.
- Have data entry staff been trained in the use of the data entry system –how long was spent on the training – do you have any mechanisms for checking competency of data entry staff?
- Are you using double data entry (DDE) – if yes, then who will carry out the data comparisons and how will this be done?

### Quality Assurance

How do you intend to ensure that your data are of high quality? Detail the data checks you are intending to carry out with a comprehensive checklist of consistency checks – for example, if a farmer only has access to one acre of land then he/she cannot be using more than one acre for growing crops; harvest date cannot be before planting date; etc.

Keep an audit trail detailing problems and inconsistencies found in the data and what was done about them. Include this information as part of the data quality assessment document.

### Data Structure & Organisation

Describe the structure of the data; in particular how many levels you are expecting and what they are; for example, village level, household level, and individual level. How many records/cases are expected at each level or is this variable. This should match with the expected number of study units from your sampling scheme.

- What field(s) will be used to link the data at the different levels?
- What formats are to be used for storing the data? Note the format for data entry may differ to that used for storage and archiving. For example, data entry might be done in CSPro but exported to SPSS for storage, analysis and archiving.

### Data Dictionary

The data dictionary should include the following information for each variable:

- **Name**– variable names should be kept short, ideally no more than 8 to 10 characters, this makes them easier to use when programming. Do you have a naming convention for your variables?
- **Label** – the label gives an indication of what the data represents. For example: name=RESPAGE, label = “Age of respondent in whole years”
- **Codes and labels** – If numeric or short text codes are used then the dictionary should detail these for each variable. For example, 1=Male, 2=Female.
- **Missing value codes** – it is useful to include codes for missing values. This should always be a value that is not feasible for the variable. For instance, a missing value code of “99” would be suitable for coded data where there are only 20 valid codes (01 to 20) but would not be suitable for “Age of respondent” as it is possible that the respondent is 99yrs of age. You might also want to distinguish between different types of missing value. For example, “not applicable”, “refused to answer”, etc.
- **Unit of measurement** – where relevant make sure you include the unit of measurement used for the data.

The data dictionary should also indicate the unique identifiers or primary key fields. Specify the number of records and the number of variables.

- Will you be deriving any variables such as a standard set of indices? If yes, describe these derived variables and explain how they are calculated. Include the syntax for creating these variables.

### Storage & Backup

Describe how you intend to make the data available to all members of the project team and how you will keep others informed about updates.

- Will you be using a DDS (Data and Document Storage Facility)? If yes, describe the system to be used.
- Will you have a file and folder naming convention? If yes, say who is responsible for ensuring this is followed?
- How will you manage the DDS to avoid accidental deletions? How will you keep it organised? Will it be password-protected? Will all team members have Read/Write access or only the data manager?
- Will you keep different versions of files? If yes, how will you distinguish between them – e.g. by including the date in the filename?

Detail your backup procedures.

- What mechanisms do you have for ensuring the security of your data?
- How often do you take backups?
- Who is responsible for backing up the data and documents?
- Which files are included in your backup?
- Where are the backups stored?
- Have you tested your restore method?

Don't rely on others to do your backups!

## Legal Aspects

### Ethics & Privacy

- Are you collecting any personal data in this research activity?
- Have you obtained ethical approval – what organisation granted this approval?
- Have all fieldworkers been adequately trained to be able to explain the consent process to potential respondents?
- Have you prepared an information sheet for respondents together with a consent form?
- What mechanisms do you have for ensuring the confidentiality of personal data?
- Detail steps taken to anonymise the data

### Data Ownership

Have all team members been made aware and agreed to the terms of the Project Data Ownership agreement?

### Copyrighted Material

Are you using any copyrighted data collection tools or methods in your research activity? If yes, have you sought permission from the copyright holder? Include here details of the copyrighted material together with any legal restrictions which might impact on how the data are used.

### Archiving & Preservation

The decision on whether to archive data is taken at the project level. The decision on where to archive might be at the project level or at the activity level.

- Where will the data be archived?
- Will there be any restrictions on access to the archive?

For further information on archiving please see the separate guide “Principles for Archiving and Sharing Data”.

### Training & Responsibilities

Name the individuals responsible for ensuring these tasks are carried out. This would normally be the data manager but might also involve scientists within the team.

Describe how you are intending to make team members aware of their responsibilities and the requirements of data management. For example, are you going to have seminars or training events or produce documented guidance?

Do you have a time scheduled for training the data entry staff? How long will this training take?

## Summary

So, to summarise there should be a policy document at the project level created by the PI which details the key data management principles. Then, for each activity, there should be a data management plan which details the steps you intend to follow to put these principles into action. This would normally be done by the project data manager. Depending on the size and scope of the activity he/she may well liaise with the team of scientists for the activity perhaps delegating from the data management tasks. Thus, we have:

- Project Level Policy – general principles of what you intend to do

- Activity Level Plan – details of how you intend to do it

## External Resources

- [IHSN – Principles and Good Practice for Preserving Data](#)
- [ICPSR – Guidelines for Effective Data Management Plans](#)

## Associate Videos

Videos accompanying the original release of the CCAFS Data Management Support Pack in 2013 are available as a playlist on the Statistical Services Centre YouTube Channel at <https://www.youtube.com/channel/UCs7EU95YMjvhNozJKCD92xQ/playlists>. These videos have not been updated since the original release but are mostly still relevant.

In particular the playlist includes a video on Data Management Plans available from the following link: [https://www.youtube.com/watch?v=Q8jX\\_cHOC60&index=3&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpi](https://www.youtube.com/watch?v=Q8jX_cHOC60&index=3&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpi)

## Reference

### 14. ToR Data Management Roles - General

- *Technical*
  - *Choice of in-house vs external services/development*
  - *Data integration across project areas/activities*
  - *IT Systems (hardware, software, services)*
- *Managerial*
  - *Data Quality Assurance*
  - *DM System Definition/understanding*
  - *Informed selection of data collection tools/methods*
  - *Team composition/skill sets needed*
- *Strategic Planning*
  - *Project level Data Management and Sharing Strategy/Plan*
- *Reference*
- *When*
  - *Decisions while designing*
  - *PI, Technician*

Please find below the document

# Terms of Reference Data Management Roles

## Introduction

In this document we have included general ToRs for a Data Manager and a Data Technician. These terms of reference are provided to help PIs and researchers with the task of determining the responsibilities of a data manager and other data management staff and as guidance for what should be expected from a data manager. These ToRs will need modifying and contextualising to make them appropriate for specific research projects. It is important to budget for these roles; see the separate document on Budgeting and Planning for Data Management for further details.

## Data Manager

The role of the data manager is to provide expertise to guide, support and oversee data management within the project.

He/she should have experience in handling data of different types including:

- Survey, experimental and monitoring data;
- Video and audio files;
- Images and photos;
- Digital maps.

## Role and Responsibilities

- Develop, maintain and enhance project databases, wherever located, while maintaining data quality standards in accordance with the project data management strategy.
- Interact with researchers within the project and beyond, and support them on issues related to data access and management, including the production of data that can be archived and published via project databases.
- Collaborate with project staff to organise and publish databases using basic web interfaces and languages, including web-based visualisation of spatial data and maps.
- Keep abreast of developments in information technology and assist in advising the project team on latest information technology solutions.
- Contribute to the development and implementation of the project data management strategy.

## The data manager should have:

- A professional degree in related fields (IT desirable);
- Comprehensive knowledge and up-to-date understanding of computing systems, databases and programming;
- Good knowledge of data management science, spatial analysis and metadata standards.

## He/she should:

- Value the sharing of information and continuous improvement in a cooperative atmosphere of constructive evaluation and learning;



- Be a team player with excellent interpersonal and communication skills, with the ability to effectively interact with people at all levels in a multi-cultural and multi-disciplinary environment;
- Have the ability to write and produce clear documentation;
- Be a completer-finisher and have the ability to notice details while keeping in mind the big picture.

## Data Technician

The main role of the Data Technician is to provide technical support to the Data Manager.

### Role and Responsibilities

The data technician would be expected to assist the data manager in a variety of roles. These could include:

- Developing data entry systems for data collection using:
  - Hand-held devices (e.g. ODK)
  - Paper questionnaires (e.g. CSPro)
- Running quality assurance checks on data;
- Extracting data for analysis;
- Writing data management guidelines;
- Helping to train project members on data management issues;
- Assisting teams with the creation and management of a data and document store (DDS);
- Assisting the data manager in implementing the project Data Management Strategy and other initiatives related to data.

### The data technician should have:

- A professional degree in related fields (IT desirable);
- Comprehensive knowledge and up-to-date understanding of computer systems, databases and programming;
- Good knowledge and experience of data management science and metadata standards.

### He/she should:

- Value the sharing of information and continuous improvement in a cooperative atmosphere of constructive evaluation and learning;
- Be a team player with good interpersonal and communication skills;
- Have excellent organisational skills and be able to demonstrate attention to detail;
- Be willing to carry out repetitive tasks and see a job through to the end;
- Be able to write and produce clear documents and guidelines for researchers;
- Have some experience in data archiving.

## 15. Example Data Management Activity Plan

- *Data Management*
  - *Archiving*
  - *Data Collection*
  - *Data Storage*
  - *Data Structures*
  - *Maintenance*
  - *Quality Control*
- *Reference*
- *When*
  - *Decisions while designing*
  - *Management of research processes*
- *Researcher*

Please find below the document

# Example Data Management Plan at Activity Level

## Introduction

This plan describes the data management activities related to the CCAFS Household Baseline Survey.

The basic data management principles related to the CCAFS programme are:

- Data generated by the project will be placed into the public domain with restricted access for confidential data such as names and GPS coordinates;
- All data generated by the project are owned by the CGIAR consortium;
- Informed consent must be obtained from respondents who must be made aware that they can withdraw from the study at any time.

## Data Collection

### Capture Methods

The aim of this activity is to gain an understanding of current practices among farmers and how these have changed in recent years. To this end we will be conducting an interviewer-led household survey.

Paper copies of the questionnaires will be completed by the interviewers. Questionnaires will be translated into the local language prior to the interviews.

### Data Description

The survey will initially be run in 15 sites spread over 12 countries from the three regions of West Africa, East Africa and South Asia. In each site a block with interesting agricultural systems and institutional links will be selected and we will list all the villages within each block. From this list we will randomly select seven villages. In each of the seven villages we will list all the households and from these lists randomly select 20 households from each village to be surveyed.

Thus, we expect to have 140 households for each site. Each site will have a unique identifier which will be a 4-character code – the code will consist of 2 letters representing the country plus 2 digits. Blocks will be numbered uniquely within the survey. Each selected village will have a unique code which is 4-characters and can be any combination of letters and digits. Within the villages the selected households will be numbered. Thus, the unique identifier for a household will be the combination of Site ID, Block ID, Village ID and Household ID.

We will collect data on the following:

- The most important crops currently grown;
- Whether this has changed in the last 10 years;
- The most important types of livestock kept by the households and whether this has changed in the last 10 years;
- What farming practices have changed in the last 10 years and reasons for these changes;
- Do these changes affect specific crops and/or livestock;
- What items are currently produced on-farm and what are produced/collected off-farm;
- What access do farmers have to land and water;
- Do they use fertiliser;

- Do they have access to information regarding climate events; e.g. information regarding the start of the rains or long and/or short-term forecasts;

We will also be asking questions regarding:

- Food security – e.g. are there times in the year when they struggle to find enough to eat;
- Other sources of income (besides farming);
- Membership of community groups;
- Assets owned by the household;
- General demographic information such as household size and composition, education level and family type (male or female-headed).

For tracking purposes, we will collect names of the household head and the main respondent – we expect these to be the same in most cases – together with GPS coordinates of the dwelling. This tracking information will not be put into the public domain and is solely for enabling us to revisit the same households in the future.

Most of the variables will be coded and a code sheet will be produced with a set of standard codes. We expect additional crop and livestock codes to be used in each site and after data collection the data manager will work on merging these crop and livestock codes to produce a comprehensive consolidated list. This will include appropriate recoding in the data files. A document will be produced detailing the merging method used and list any recoding that will have been done to the data.

To record the amount of land that households have access to we will record and use the local unit of measurement. The supervisor will then record the conversion factor to convert these values to hectares.

## **Computerisation & Storage**

### **Data Entry**

Data will be recorded on paper questionnaires which will be visually checked for completeness and consistency by both the enumerators and the supervisors. Both will sign and date the front of the questionnaire when they have done these checks. A checklist will be produced to ensure that the same checks are done on each completed questionnaire.

For data entry we will be using CPro. A consultant will produce the data entry system with screens that resemble the questionnaire as far as possible. The data entry system will be programmed to follow the skips present in the questionnaire. A data entry manual will be produced to accompany the system, and this will be used when training the data entry staff. Once the questionnaire is finalised a list will be drawn up of possible consistency and range checks that can be programmed into the data entry system. The documentation will detail these checks.

A version of the data entry system with the screen labels in French will be produced for use in West Africa.

The data entry system will be thoroughly checked by entering data from 5 completed questionnaires and adjustments made as appropriate.

Data entry staff will receive 2 or 3 days of training on using the system. They will each receive a copy of the manual and a log book to record problems.

## Quality Assurance

At the end of each day the field supervisors will visually check each completed questionnaire for obvious errors. Instructions for these checks will be included in the Field Supervisors Manual.

Double data entry will be used. For each site two data entry clerks will each enter data from all 140 questionnaires into separate data files. The data entry supervisor or data manager will then use the CSPro Data Compare tool to look for differences in these files. Any differences will be checked against the original questionnaires and corrected in both data files. The output report from the Data Compare tool will be stored with the project archive as part of the audit trail of data quality checks.

After the double data entry comparisons further checks on the data will be done. Frequency tables will be produced on all coded variables to ensure that there are no values outside the expected range. Frequency tables on the villages will help us check that we have the correct number of households per village. Names and GPS coordinates will be checked against the sampling frame files to make sure we have collected data from the sampled households.

For the land values we will do the following:

- Check that the total amount of land owned is always greater than the amount of owned land used for grazing, growing crops, etc.
- Do the same for rented land – i.e. the amount of rented land used for grazing cannot be greater than the total amount of rented land.
- Look at the highest and lowest values to ensure they are reasonable and that no one farmer has an order of magnitude more land than any other farmer. We will produce histograms or boxplots of land values so that we can easily see the range of data. Any extreme values will be investigated.
- The data will be compared against local knowledge – for example, we might know that in a particular site, farmers tend to have very small plots of land, so if the data are showing very large farms then we will investigate. We need to ensure for example that the conversion factor is recorded correctly. The conversion factor would be the number of land units in a hectare.

A log will be kept of potential errors and outliers in the data, together with a report of how we dealt with these values.

Once the data are checked and corrected a data quality assessment document will be produced.

## Data Structure & Organisation

The main study unit will be the household and we expect 140 household level records per site. There may also be data at other levels – for example plot level, individual level, etc. Until the questionnaire is finalised we will not know what other levels there will be. The data entry system will be set up so that data from all levels will be entered at the same time into the same CSPro file. Once exported, data from each level will be stored in separate data files and each will include the primary key fields from the household level, namely Site ID, Block ID, Village ID and Household ID. These fields will act as the link between data at different levels.

The data will be entered using CSPro but once the data comparisons and data checks are completed, they will be transferred to SPSS. There will be one CSPro data file but separate SPSS files for each level in the data. Syntax files will be produced for labelling the data and for calculating some standard

indices. As yet the set of standard indices have not been decided but these will be documented when ready.

### Data Dictionary

A data dictionary will be produced containing the following information for each variable:

- **Name**– variable names will be no more than 8 characters in length and the names will be printed on the final questionnaire so that researchers can easily see which question a variable refers to.
- **Label** – the label will be the question text or an abbreviation of it.
- **Codes and labels** – any numeric codes used will be listed. A code book will be produced for cases where several variables use the same set of codes and the dictionary for the variables will refer to the code book.
- **Missing value codes** – we will use three missing value codes throughout
  - **-8** will be used to indicate “Not applicable”
  - **-6** will indicate no consensus among family members
  - **-9** will be for any other missing value
- **Unit of measurement** – the unit of measurement for amounts of land will vary between sites; therefore, there will be a variable containing the name of the local unit and a second variable containing the conversion factor to convert the local units to hectares.

The primary key will be the combination of Site ID, Block ID, Village ID and Household ID. The set of derived variables has not yet been decided.

### Storage & Sharing

The data manager will be responsible for producing guidelines for project members on the structure of the shared folders – the DDS. The guidelines will include how to name files and the expected times and modes of delivery for project files.

We will be using Dropbox to share project files among team members. One Dropbox share will have read/write access to all team members but there will be a separate share to which only the data manager has write access and others are included by way of links. In this share the data manager will start to build the archive and will inform team members of any updates and additions via email. All files in the read-only share will include dates in the file name – dates will be in the form YYYYMMDD – e.g. HouseholdQuestionnaire20110516.docx

Files on Dropbox will be backed up automatically, but the data manager will also make daily backups of the Dropbox folders. These will be stored on the DMs own PC.

### Legal Aspects

#### Ethics & Privacy

An information sheet will be prepared for respondents which describes the study. This, and the consent form, will clearly state that the data collected will be used for research purposes and that data will be made public although any identifying information will be removed prior to archiving. The box below shows the consent statement that will appear at the top of the questionnaire.

*“Good morning/afternoon. We are coming from (\_partner organisation’s name\_) with permission from the local government. We are conducting a survey **looking at farming practices and how***

**they change over time.** We would like to ask you some questions that should take no more than one to one and half hours of your time. We would like to share some of this information widely in order that more people understand how food is grown and used in this region and the issues that you face regarding food production and soil, water and land management.

Your name will not appear in any data that are made publicly available. The information you provide will be used purely for research purposes; your answers will not affect any benefits or subsidies you may receive now or in the future. Do you consent to be part of this study? You may withdraw from the study at any time and if there are questions that you would prefer not to answer then we respect your right not to answer them.

Prior to archiving, names and GPS coordinates will be removed from the main data files. Only village codes will be used in the data files rather than village names. Separate files containing the names will have restricted access in the archive.

### Data Ownership

All team members are aware of the data ownership agreements of the project and have signed the agreement. Any new team members will be asked to sign the agreement before being able to access the data.

### Copyrighted Material

Digital maps used are copyrighted.

### Archiving & Preservation

A Dataverse will be created for the Baseline studies and data and documentation will be uploaded. The original CSPro data files will be loaded into the Dataverse but will have restricted access. The archive will include final versions of all files from the household survey up to and including site reports. The following files will be included:

- Questionnaires – in all languages
- Code book
- Fieldworker manual
- Data entry manual
- Data checking guide
- Data entry system
- CSPro data files – restricted access
- Analysis plan
- SPSS syntax files for labelling data, calculating indices and carrying out standard analyses
- Output from analysis plan
- Merged and anonymised data files in SPSS format
- Household Identification information – restricted access
- Sampling Frames – restricted access
- Site analysis reports
- Process reports – e.g. data quality assessment document

The Dataverse will be set up within 24 months of the end of data collection.

## **Training & Responsibilities**

Within each country team an individual will be given data management responsibilities and will report to the overall project data manager. The data manager will have overall responsibility for ensuring the files are submitted in a timely manner.

Data entry staff will receive training on using the data entry system. Data comparisons will be done by the local team if there is anyone in the team with the required skills. Otherwise the files will be sent to the project data manager for the comparisons to be done.



## 16. Tools for Research Projects

- *Data Management*
  - *Data Collection*
  - *Maintenance*
  - *Quality Control*
  - *Data Storage*
  - *Archiving*
- *Technical*
  - *IT Systems (hardware, software, services)*
  - *Choice of in-house vs external services/development*
  - *Data Security*
  - *Data Integration across project areas/activities*
- *Using the Data*
  - *Primary User – Researcher*
  - *Understanding the DM Process*
  - *Dashboards/data displays*
- *Reference*
- *When*
  - *Decisions while designing*
  - *Management of research processes*
  - *Delivery of research products*
- *Researcher, Technician*

Please find below the document

# Tools for Research Projects

## Introduction

In this document we look at some of the tools that a researcher might use during the lifetime of a project. We have included a brief paragraph about each tool mentioned and have included web links where researchers can find further information. This document is intended to be a “live” document which should be regularly updated as new tools become available; web links should also be checked periodically to ensure they are still valid.

This document is not intended to be a list of recommendations; we do not promote any particular software solution but simply wish to share information about tools that we have used and have found useful for research projects.

## Data Collection Tools

In this section we look at a few tools that can be used for data collection and data entry. Some of these are specifically used for data collection on hand-held devices; e.g. ODK; while others, such as CSPro can also be used for entering data from paper questionnaires.

### Open Data Kit (ODK)

ODK is a free and open-source set of tools for building data collection forms for Android devices. Forms are generally built using the XLSForm standard, which was created to help simplify the authoring of forms using Excel. It is relatively easy to get started on using ODK and as you become more proficient you can build up to quite complex forms including skip logic, data checks, and repeating groups. For information on writing data entry forms for ODK, see <http://xlsform.org/>.

Once the form is complete it is converted to a format that can be installed on handheld devices. ODK Collect is then used on the device to collect the data during the interview. Once the data have been collected and checked, they can be uploaded to an aggregate server from where they can be analysed or exported. This process allows for monitoring of data collection, and progress reports can be quickly and easily produced during the fieldwork.

ODK is typically used with a server, which acts as the central point for uploading your forms and aggregating collected data from multiple devices and users. There are many server options available to projects, including third-party services (both free / open and commercial) and self-hosting options.

Your organisation’s IT / Data Management supporters will likely have recommended services to use, and your organisation may even have their own ODK Aggregate server that you can use for your project. We recommend discussing your requirements within your organisation to learn if such services are available.

For further information on ODK go to <https://opendatakit.org/>

### CSPro

The Census and Survey Processing System (CSPro) is a public domain software package for entering and managing census and survey data. It is developed and supported by the US Census Bureau. CSPro

is designed to be as user-friendly as possible yet is powerful enough to handle complex applications. The software runs on Windows PCs but there is also an CSEntry Android App which works in collaboration with the desktop version of CSPro. Data entry applications would be created under Windows and compiled versions would be transferred to the Android device for data collection. Data files can then be uploaded to Dropbox or to an FTP server while in the field.

Alternatively, data from paper questionnaires can be entered directly into the Windows application using forms that can be designed to resemble the questionnaire.

For further information on CSPro go to <https://www.census.gov/data/software/cspro.Overview.html>

## Surveybe

Surveybe is a computer assisted personal interviewing (CAPI) software suite. The suite comprises two elements: the Designer and the Implementer.

The Designer is used to build, configure and update a questionnaire, including screen structure, questions, rosters and validations. The Implementer is then used to display the questionnaire on the device and collect the data.

Note: Surveybe is not a free product; you would need to buy a license for the Designer. However, when you buy a Designer license, the Implementer is free. Thus, with a single Designer license you can do all your designing on a single PC but use the Implementer on any number of devices for data collection.

For further information on Surveybe, including pricing, go to <http://surveybe.com/>

## KoBoToolbox

KoBoToolbox is a suite of tools that uses ODK for field data collection. The main service is the Kobotools server, which fulfils the role of the ODK Aggregate server. This service is freely provided, as it is funded by the Harvard Humanitarian Initiative and a group of other international organisations.

Kobotools also has their own data collection application, KoboCollect, which is essentially a clone of ODK Collect. KoboCollect is more visually compatible with Kobotools, but ODK Collect is generally more up-to-date and receives new features and bug-fixes at a faster rate than Kobotools.

Kobotoolbox also provides an online form builder, which is useful for building quick data collection forms without needing to write the XLSForm yourself. For more powerful forms, it is usually easier to author the form in XLSForm, which is fully compatible with the Kobotools server.

For further information visit the website at <http://www.kobotoolbox.org/>

## Survey Solutions

Survey solutions is a suite of tools developed by the World Bank to support CAPI / CAWI / mobile data collection projects. It includes an online questionnaire designer, Android data collection application and “headquarters” server software, that enables project managers to oversee teams of enumerators.

Most of the tools are free to use, but the systems are not open source, so you will need to contact the Survey Solutions team to find out what’s available for your project. While the World Bank resolve to offer the software for free, they may ask projects to pay for server costs, depending on the type of project.

For more information, see the Survey Solutions support portal: <http://support.mysurvey.solutions/>.

## Options for Data and Document Storage

A Data and Document Store (DDS) is primarily for sharing project files within the team during the lifetime of the project. In this section we give a few common examples of tools that can be used for a DDS. See the separate document on DDS for more information.

### OneDrive

OneDrive is a file-hosting service operated by Microsoft as part of its suite of online services. It allows users to store files in the cloud. Files can be synced to a PC and accessed from a web browser or a mobile device as well as shared publicly or with specific individuals.

OneDrive offers 5Gb of storage space free of charge; additional storage can be added either separately or through subscriptions to other Microsoft services including Office 365.

Like many syncing services, you can organise files within your OneDrive account into folders in the same way as you do in Windows on your local computer. Microsoft offers a syncing service for Windows and Mac, which lets you sync your OneDrive files to your local drive, allowing you to work with files offline. They also offer apps for iOS and Android to access your stored files.

On Windows 10, OneDrive offers selective syncing, which lets you view all your OneDrive files in Windows Explorer (as if they were stored locally), but to choose only certain files to have downloaded. This lets you view and access all your files without taking up space on your local drive. This is very useful if you have a large amount of content in OneDrive and a small disk in your local computer, but it's important to remember to download key files if you are going to be offline for an extended period.

For further information about OneDrive, visit the website at <https://onedrive.live.com/about/en-gb/>

### Dropbox

Dropbox is an easy to use and very popular file hosting service that gives access to files through the web. It enables easy sharing of files both as full access shares or through read-only links.

Like OneDrive, Dropbox offers Windows and Mac apps to sync your files to your local drive. The app creates a special folder on the user's computer and the contents of the folder are then synchronized to Dropbox's servers and to other computers and devices that the user has installed Dropbox on. Users can invite others to share one or more of their folders.

Dropbox Basic users are given 2Gb of free storage, but you can "earn" additional free space in a number of ways (see the web site for more details). Dropbox Plus and Dropbox Professional are paid subscriptions that both include 1Tb of space plus some additional features. Finally, there is Dropbox Business which is geared towards organisations and groups. Pricing depends on the size of your team and your billing country.

For further information on Dropbox see the website at <https://www.dropbox.com/> - there are also many introductory videos available on YouTube.

## Google Drive

Google Drive is a file storage and synchronisation service. Google Drive offers 15Gb of free storage, with up to 30Tb offered through paid plans. Google Drive is a key component of G Suite, Google's monthly subscription offering for businesses and organisations.

As with the other cloud storage options, Google Drive offers Windows and Mac apps to sync your files to a folder on your local drive. The same app also allows you to backup any other folders on your local drive, but this also takes up space in your Google Drive account.

For further information, see the website at <https://www.google.com/drive/> - again there are many introductory videos available on the web.

Google drive also links to Google's online office suite – Docs, Sheets and Slides. For information on these services, see <https://www.google.com/docs/about/>

## Database Systems

Some projects' requirements for data storage and management can be addressed by careful management of data files, such as csv, Excel workbooks or json files. For project with more complex data structures, or projects that need to manage an evolving dataset over a period of time, you will likely require a database system to help manage your data.

The biggest decision to make is whether to run your database on your local computer or on a server.

- Local installations are easier to setup and have fewer dependencies. They also allow you to use your data whenever you need without having to worry about having a stable internet connection. However, only you have access to the database, and anyone else wanting access must take a copy, and changes made in different locations will not be automatically synchronised. You must also ensure you have adequate backups in place, in the same way as you would for regular files on your local disk.
- Databases running on a server are more complex to setup, and it is likely you will need the support of an IT / network expert to manage the system. Your organisation may already have infrastructure in place for one or more of the database types you require, so if you want to make use of a database on a server, talk to your IT team. They will also be able to advise on the best technical solutions for your project needs.

Advantages of a database running on a remote server include:

- It can act as a central location for your "truth" data, allowing your entire team to work with the same dataset.
- All good systems have options for managing access levels, so you can choose what data to share with different users, and what permissions they have with those data.
- Remote servers generally have redundant hardware setups, so you are not relying on a single piece of hardware (e.g. your local hard drive).
- A good database setup will also have a way to record changes to your data, so you can maintain a unified record of all the changes that any member of your team makes during the course of the project.

The rest of this section lists different database packages that are available. It is not an exhaustive list, but covers many of the popular options.

## MS Access

Microsoft Access is a powerful database management tool that allows a data manager to setup customised data entry forms to allow non-technical users to enter data and interact with the database effectively.

It is a good option if you require a database that can run locally, or shared among users on local network storage. Newer versions of MS Access also offer the option of creating cloud-based databases and applications, but these are generally less versatile than other solutions for cloud-based databases.

For more information, see the Microsoft site: <https://products.office.com/en-gb/access>

## MySQL

MySQL is a free, open source SQL (structured query language) based database maintained by Oracle. It is one of the most flexible SQL database systems available and is easy to start using if you are new to structured databases.

Oracle provides installers for Windows, Mac and Linux to install MySQL locally: (<https://dev.mysql.com/downloads/installer/>), but the main way of using MySQL is to install it on a remote server that your project has access to.

MySQL is a good option to choose if you have a clear data structure with well-defined relationships between data objects and levels. Of the “SQL” based options, it is the easiest to get setup if you do not have much experience with database management.

## PostgreSQL

PostgreSQL is another popular open source SQL-based language. It is similar to MySQL and shares many of the same traits, including the support for defined structures and data relationships. (<https://www.postgresql.org/>).

In some ways, it is far more powerful than MySQL, and is a good option to choose if you are working with extremely large structured datasets. It also has a powerful suite of tools for handling geographic objects and running location-based queries called PostGIS (<https://postgis.net/>). It is also considered more complex than MySQL and is not as easy to setup if you are unfamiliar with the options available.

## MSSQL Server

MS-SQL is Microsoft’s commercial SQL-based database. It is typically used in Windows Server / Dot-net environments and runs well in those situations. It is targeted more towards business users and teams already embedded into the Microsoft environment.

For more information, see the Microsoft website: <https://www.microsoft.com/en-gb/sql-server/sql-server-2016>

## MongoDB

MongoDB is an open source noSQL database, developed and maintained by MongoDB Inc. (<https://www.mongodb.com/>). It is a popular database system for developers and data analysts, as it is highly scalable and performs well with extremely complex queries. Data are stored as JSON objects, which can be grouped into collections for easier management, and nested to account for multi-level datasets.

The database software itself is open source, but MongoDB Inc have a wide range of commercial offerings, including different levels of hosting and management.

For more information about the use of MongoDB, see their documentation site: <https://docs.mongodb.com/getting-started/shell/tutorial/install-mongodb-on-windows/>

## CouchDB

CouchDB is another JSON-based noSQL database, maintained by the Apache group. (<http://couchdb.apache.org/>).

CouchDb provides a powerful API to interact with the database programmatically, and offers data replication options to allow users of applications to synchronise a dataset and continue to work offline.

## Analysis Software

In this next section we look at a few of the more popular statistical analysis software packages.

### R

R is a free, open source software environment for statistical computing and graphics. It runs on Windows, MacOS and UNIX platforms.

R provides a wide variety of statistical and graphical techniques and is highly extensible. One of its strengths is the ease with which well-designed publication quality plots can be produced.

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License. R can easily be extended via "packages". There are about eight packages supplied with the R distribution and many more are available.

At the time of writing the latest version of R is version 3.4.3 which was released on 2017-11-30. It is a good idea to visit the website regularly for new releases.

For further details visit the website at <https://www.r-project.org/>

### RStudio

RStudio is a front-end UI for running R. One of the challenges of using R is the basic nature of the interface and the challenges of needing to write commands for everything. RStudio is a popular option that gives UI controls for common tasks like importing datasets, loading package libraries and reviewing scripts.

For more information, see <https://www.rstudio.com/products/RStudio/#Desktop>

### Stata

Stata is a general-purpose statistical software package. Its capabilities include data management, statistical analysis, graphics, simulations, regression and custom programming.

Stata licenses are generally for perpetual use although there is now the option to have an annual subscription. A single-user license can be installed on up to three of your personal computers as long as you are the sole user. See the website for pricing options.

At the time of writing, the latest version of Stata is version 15 (June 2017)

For further information about Stata visit the website at <https://www.stata.com/>

## SPSS

SPSS (officially named IBM SPSS Statistics) is a widely used program for statistical analysis particularly in the social sciences. SPSS is reasonably easy to use either with the dropdown menus or via syntax files. SPSS is not cheap and there is an annual license fee. However, there is a GradPack available for current students which gives a substantial saving (up to 99%) on the standard annual license fee.

For further information about SPSS, visit the website at <https://www.ibm.com/products/spss-statistics>

## GenStat

GenStat is a statistical software package with data analysis capabilities particularly in the field of agriculture. It is developed and marketed by VSN International Ltd (VSNi). GenStat licenses work on an annual subscription; you would need to contact VSNi directly for pricing information. Students can purchase the full version of GenStat at a reduced rate; the reduced rate is an annual fee and is only available to users in educational (degree granting) organisations.

For further information about GenStat, visit the website at <https://www.vsn.co.uk/>

## Archiving Options / Repositories

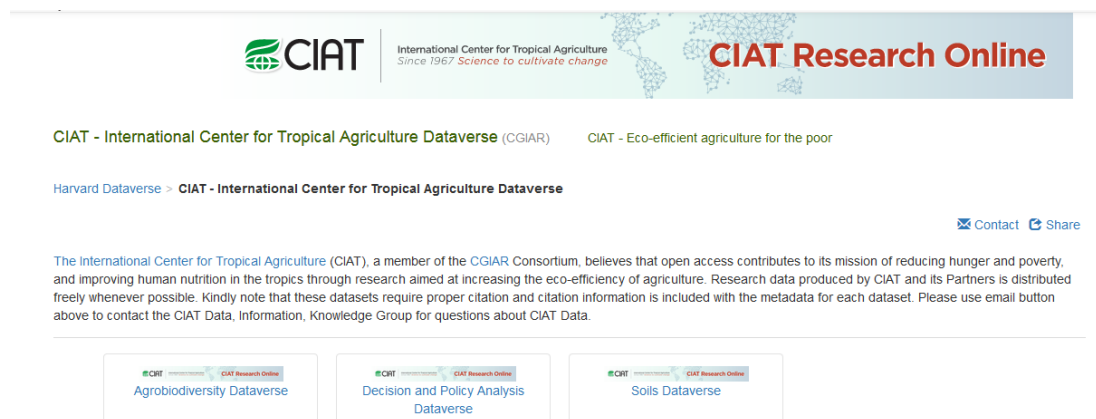
In this section we briefly mention a couple of the more popular data and document repositories often used by researchers in the CGIAR.

### Dataverse

Dataverse is an open source web application for sharing, preserving, citing, exploring and analysing research data. It facilitates making data available to others and allows you to replicate others' work more easily. Researchers, data authors, publishers, etc., all receive academic credit and web visibility.

A Dataverse repository hosts multiple dataverses which in turn contain datasets. Each dataset contains descriptive metadata and data files (including documentation and code to accompany the data). A separate document in this pack describes Dataverse in more detail.

CIAT and CCAFS both have their own dataverse and the image below shows the CIAT dataverse which contains three sub-dataverses: Agrobiodiversity, Decision and Policy Analysis, and Soils.



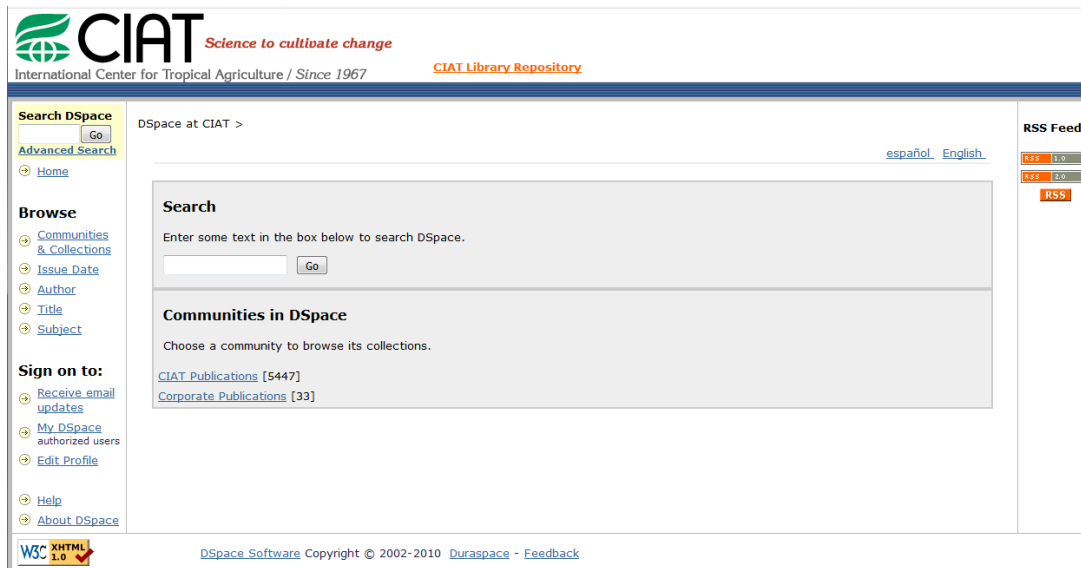
The screenshot displays the CIAT Research Online website. At the top, the CIAT logo is on the left, with the text "International Center for Tropical Agriculture Since 1967 Science to cultivate change" in the center, and "CIAT Research Online" on the right. Below this, there are two main sections: "CIAT - International Center for Tropical Agriculture Dataverse (CGIAR)" and "CIAT - Eco-efficient agriculture for the poor". A breadcrumb trail shows "Harvard Dataverse > CIAT - International Center for Tropical Agriculture Dataverse". There are "Contact" and "Share" buttons. A paragraph of text describes CIAT's mission and data policy. At the bottom, three sub-dataverse boxes are shown: "Agrobiodiversity Dataverse", "Decision and Policy Analysis Dataverse", and "Soils Dataverse".

For more information about Dataverse, visit the website at <https://dataverse.org/>



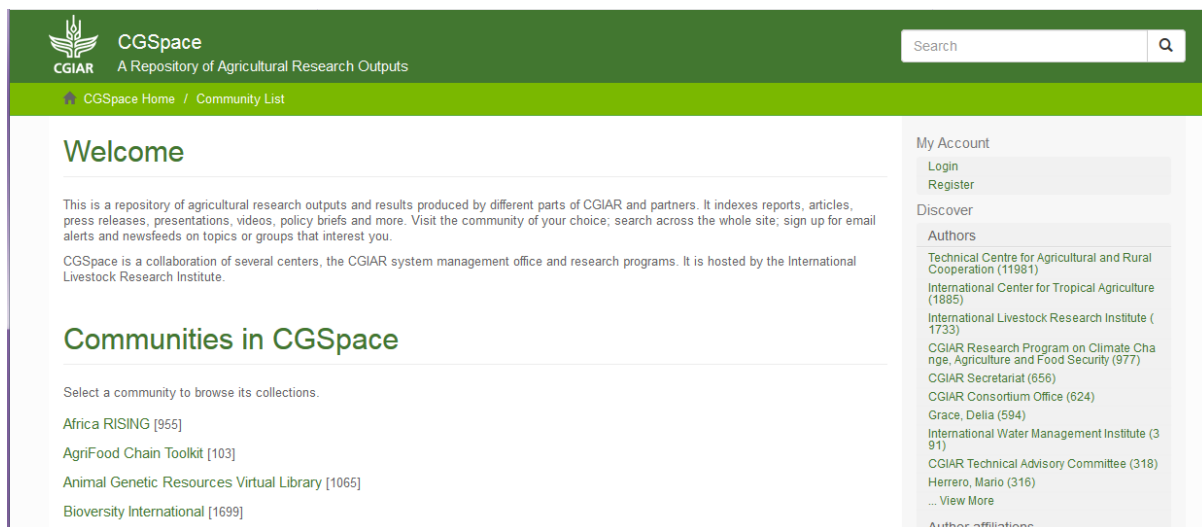
## DSpace

DSpace is software for building open digital repositories. It is free and easy to install and can be customised to fit the needs of any organisation. CIAT has its own Library Repository using DSpace which is available at <http://ciat-library.ciat.cgiar.org:8080/jspui/>



The screenshot shows the CIAT DSpace Library Repository interface. At the top, the CIAT logo is displayed with the tagline "Science to cultivate change" and the text "International Center for Tropical Agriculture / Since 1967". Below the logo, the text "CIAT Library Repository" is visible. The main content area is titled "DSpace at CIAT >" and features a search bar with a "Go" button. To the left of the search bar, there are navigation links for "Home", "Browse", and "Sign on to:". The "Browse" section includes links for "Communities & Collections", "Issue Date", "Author", "Title", and "Subject". The "Sign on to:" section includes links for "Receive email updates", "My DSpace authorized users", and "Edit Profile". Below the search bar, there is a section titled "Communities in DSpace" with a list of communities: "CIAT Publications [5447]" and "Corporate Publications [33]". On the right side of the interface, there are "RSS Feeds" for "RSS 1.0" and "RSS 2.0". At the bottom of the page, there is a "W3C XHTML 1.0" logo and a copyright notice: "DSpace Software Copyright © 2002-2010 Duraspace - Feedback".

In addition, the CGIAR have set up CGSpace at <https://cgspace.cgiar.org/> This is a repository of agricultural research outputs and results produced by CGIAR centres, initiatives and challenge programmes including CCAFS. The contents of the site can be browsed according to region, author, CGIAR centre, programme, etc.



The screenshot shows the CGSpace interface. At the top, the CGIAR logo is displayed with the text "CGSpace A Repository of Agricultural Research Outputs". Below the logo, there is a search bar with a "Search" button. The main content area is titled "Welcome" and contains a paragraph of text: "This 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. Visit the community of your choice; search across the whole site; sign up for email alerts and newsfeeds on topics or groups that interest you." Below this text, there is a section titled "Communities in CGSpace" with a list of communities: "Africa RISING [955]", "AgriFood Chain Toolkit [103]", "Animal Genetic Resources Virtual Library [1065]", and "Biodiversity International [1699]". On the right side of the interface, there is a "My Account" section with links for "Login" and "Register". Below this, there is a "Discover" section with a list of authors: "Technical Centre for Agricultural and Rural Cooperation (11981)", "International Center for Tropical Agriculture (1885)", "International Livestock Research Institute (1733)", "CGIAR Research Program on Climate Change, Agriculture and Food Security (377)", "CGIAR Secretariat (656)", "CGIAR Consortium Office (624)", "Grace, Delia (594)", "International Water Management Institute (391)", "CGIAR Technical Advisory Committee (318)", "Herrero, Mario (316)", and "... View More". At the bottom of the right side, there is a link for "Author affiliations".

## AgTrials

This is the Global Agricultural Trial Repository which is an information portal developed by CCAFS which provides access to a database of the performance of agricultural technologies at sites across the developing world. With the interface you can:

- Share data and information on evaluations of agricultural technology
- Acquire agricultural evaluation datasets for your own research
- Explore the geographical dimensions of agricultural evaluation.

The repository is available at <http://agtrials.org/>

Search Trials

Simple and Advanced Searches

Add New Trial

Upload your data to AgTrials

Upload Batch of Trials

Upload your data in batch mode

### What is AgTrials?

Agtrials.org is an information portal developed by the CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS) which provides access to a database on the performance of agricultural technologies at sites across the developing world. It builds on decades of evaluation trials, mostly of varieties, but includes any agricultural technology for developing world farmers.



### What you can do

Share data and information on evaluations of agricultural technology.


Acquire agricultural evaluation data sets for your own research.

Explore the geographic dimensions of agricultural evaluation.

## CCAFS-Climate

The CCAFS-Climate data portal provides global and regional high-resolution climate datasets that serve as a basis for assessing the climate change impacts and adaptation in a variety of fields including biodiversity, agricultural and livestock production, and ecosystem services and hydrology.

The Climate portal is available at <http://www.ccafs-climate.org/>



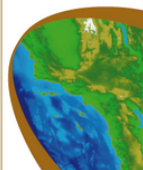
**GCM**  
DOWNSCALED  
DATA PORTAL

Contact About Us


CGIAR RESEARCH PROGRAM ON  
Climate Change,  
Agriculture and  
Food Security CCAFS

[Home](#) [Data](#) [Methods](#) [Documentation](#) [Links](#) [Citations](#)


**Data**




**Methods**




**Useful Documents**




**Links**



**Citations**



**Contact**




**Data Provided by the CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS)**

The data distributed here are in ARC GRID, and ARC ASCII format, in decimal degrees and datum WGS84. CCAFS and its partners have processed this data to provide seamless continuous future climate surfaces. Users are prohibited from any commercial, non-free resale, or redistribution without explicit written permission from CCAFS or the data-developing institutions. Users should acknowledge CCAFS as the source used in the creation of any reports, publications, new data sets, derived products, or services resulting from the use of this data set. For commercial access to the data, send requests to **Andy Jarvis** at the International Center for Tropical Agriculture (CIAT).

These open-access datasets are hosted by **Amazon Web Services**.

CCAFS provides these data without any warranty of any kind whatsoever, either express or implied, including warranties of merchantability and fitness for a particular purpose. CCAFS shall not be liable for incidental, consequential, or special damages arising out of the use of any data published here.



This work by <http://ccaafs-climate.org/> is licensed under a **Creative Commons Attribution-NonCommercial 4.0 International License**

# Data and Document Store

## Main

### 17. Creating and Using a Data and Document Storage Facility

- *Communications*
  - *Efficient sharing of information*
  - *Internal communications*
- *Data Management*
  - *Data Storage*
- *Using the Data*
  - *Sharing and Access internally*
- *Main*
- *When*
  - *Management of research processes*
- *PI, Researcher, Technician*

Please find below the document

# Creating and Using a Data and Document Storage Facility (DDS)

## Introduction

A Data and Document Storage Facility (DDS) is basically an area where project data and documents are stored, together with the rules that enable the project team to use it effectively and efficiently. For a single researcher working on his/her own, this could simply be a set of folders on his/her own PC hard drive. For a team of researchers this might be shared folders on a network file server or a web-based cloud storage system.

## History

Before the advent of the Internet and computer networks, researchers would store their data and documents on standalone PCs. This worked fine for small projects with one researcher, but as projects became larger and teams developed, you often had the situation with each team member holding some but not all of the documents and data for the project. You often ended up with duplication and various copies of files with no one really knowing which was the most up to date. In short, there was a mess and archiving at the end of the project was just too daunting a task so wasn't done.

## DDS vs Long Term Storage (Archives)

In these guides, we distinguish between 2 types of storage. A DDS is intended for active projects – where team members need regular access to data, and documents get updated regularly. Archiving occurs at the end of the project, when regular access is no longer needed, and files are not going to change much.

This may seem arbitrary, but the requirements of these 2 types of data storage are actually quite different, so it's likely you'll use different systems for them.

## Why might I need a DDS?

The fact is, you always need a DDS for your project. Whether you are working alone or in a group, you need somewhere to store your project documents and data while working on them. The main questions are about how you organise this space, and whether you need a centralised space for your team to share resources.

When working alone, it's very tempting to not put effort into maintaining an organised system. We are all guilty at times of this: giving your vital documents a 'temporary' name; dragging it to your desktop for quick access and then forgetting to file it at the end of the day. Even when you can remember the file names and locations of things in the short term, this method will eventually fall down when you realise you need a specific document 6 months later and cannot remember if your most recent analysis script was "analysis working.R" or "analysis with metadata.R". (And that's assuming you know what project the files were for!)

When working in a team, good file organisation goes from important to vital. This is where you should consider a central storage space. Instead of the data and documentation being divided among the researchers to manage individually, you create a space to ensure the data and documentation relating to the project are up to date, complete and available to those who should have access. You should

agree as a team how to use the space, and these agreements should be part of your Data Management Plan.

Having a central space doesn't stop a researcher taking a copy of part of the data to work on separately, as long as they know that any updates must be made to the master files in the DDS. The responsibility for negotiating and setting the rules for the DDS falls to the person with overall responsibility for the project, generally the PI. The responsibility for implementing the DDS is often delegated to the data manager.

## Requirements of a good DDS

You can use many different systems to create your DDS, which we discuss below. Regardless of what tools you use, you should consider the following functions and how important they are for your project.

### Storage

The primary use of a DDS is to store your active project files. This means you need to ensure you have enough space to store everything. When planning, try to predict how much storage you will need, and then set up your DDS with more than you think you need. That might mean buying more space on your cloud service account or buying extra drives for your local storage system. In 2018, storage is relatively cheap, so buying more than you think you'll need won't be a huge drain on your budget.

Beyond judgement of storage space, you may also want to consider the types of data and documents you'll be storing, and how they'll be used during the project. For example, if you are collecting large images and videos, you may want to provide local storage for people working with those files, as accessing them remotely might put strain on your network.

### Syncing

Cloud storage systems like OneDrive, Dropbox etc. usually allow users to synchronise files to their local computers. This is hugely beneficial as it allows users to access files while offline and more easily work with locally installed applications.

Some systems allow for users to choose what files get synchronised, so they can have greater control over what files get stored on their local drives. This is increasingly important as many people now use laptops with relatively small solid-state drives.

### Permissions Levels

When structuring your DDS, consider the purpose of each section, and who needs access. It is good practice to only grant specific types of access to those that need it. All good systems should let you choose between "read-only" and "read-write" for users. Beyond that, you might have other options, for example giving users admin permissions or the ability to grant access to other users. For example, when creating a shared folder in OneDrive, you can let users edit (read-write) or view (read-only). Someone with read-only access will not be able to edit the document directly within your DDS. They might be able to copy it somewhere else to edit, but your original file will remain intact and unchanged.

## Version control

Many people use a “DIY” system of version control, by saving different versions of a file with slightly different names (see “Naming conventions” above). This is hugely preferable to having just a single file that you save over each time, as it makes you less vulnerable to accidental file corruption. If you choose to use this simple method of version control, we recommend making a copy of any file **before** starting your work for the day. That way, there is no way of accidentally overwriting your file with changes that you might later want to reverse.

This simple version control is usually enough for non-vital documents and personal files, but it relies on the user manually making copies to be a good option for mission-critical documentation or vital / unique datasets. For these more important files, we recommend a more automated option.

Certain apps have a form of version control or change tracking built in. Microsoft Office applications have a “track changes” feature that lets you follow the changes different authors make to a document. Google Docs also lets you review all the changes made by every author in a timeline, and lets you restore previous versions of a document.

But your DDS system can also include version control. For example, the Dropbox business plan lets you view and restore every version of a file saved within the last 120 days. OneDrive keeps file versions for 30 days, and OneDrive for Business lets you specify a certain number of versions to keep, rather than it being based on time.

A DDS kept on a local network drive may also have the capacity to keep different versions of files. Even on a local DDS, you might be able to setup some form of automatic version control. Windows 10 has a “File History” feature, which allows you to backup versions of your files in specific folders to an external drive, then access these older versions directly within the File Explorer.

## Advanced Version Control – Git, SVN, Mercurial

There are more advanced version control systems focused on code development. Git, Subversion (SVN) and Mercurial are all systems focused around providing version control and change tracking for coding projects. They are extremely useful for software development and recommended for any project involving the writing of a lot of text files, for example doing analysis work or software development.

These systems take more time to learn and setup than a OneDrive folder or shared network drive but can provide huge benefits for certain types of work. If you want to be able to track line-by-line changes to code files, have a complete revision history (including comments) kept for you as you work, or allow different team members to work concurrently on the same codebase, we highly recommend exploring one of these systems.

Even if you’re not tracking text files, a system like Git might still work for you. There are projects that use Git to version control Word documents, Excel files and other large files (like repositories of images).

## Backups

This is important: version control is not a backup. Using OneDrive as your DDS and syncing your OneDrive folder to your local computer does not count as having 2 copies of those files, because if your local copy changes, the copy on the OneDrive server also changes. If someone accidentally deletes it in one place, it’s gone everywhere. While you might be able to use OneDrive’s version control to restore that file, that relies on a non-corrupt version being available in the system.

For this reason, your DDS plan should be more expansive than simply “We’ll use OneDrive” or “We’ll have a shared folder on the local network”. You should consider what backup processes need to be implemented to ensure your data are safe. It may be that your DDS is already covered by an existing process. For example, your entire local network storage might be backed up every night by the network administrator, which would include a snapshot of your DDS. In this case, make sure you understand where those backups are kept, how secure they are and what processes are in place to access them if needed.

In general, you should consider the “3-2-1” rule of backups: Keep at least 3 copies of your data; 2 of which are local but on different devices, and at least 1 copy offsite. (If you are restricted from truly “offsite” storage by data policies or network restrictions, then try to spread the physical media out as far as possible – ideally in different buildings). This policy makes it extremely unlikely that you’ll lose all copies of your data in one go.

A standard “manual” backup process might involve taking daily snapshots of your DDS and placing these, as compressed (.zip or .tar.gz files) on a secure storage – e.g. an external drive. To save space, you might only keep daily backups for 2 weeks. You might then keep weekly backups for up to 2 months, then beyond 2 months you might keep 1 backup for each month until the end of the project.

In practice, we recommend making the backup process as automatic as possible. The less you must rely on a member of your team manually copying files, the better. On a local Windows system, you might use the Windows Task Scheduler to run a batch file that clones your project folders to an external drive. Or you might choose to use a third-party backup tool. On a network, you might use rsync to run automatic, incremental backups of your network drive.

## Security

You need to ensure your DDS is secure from unwanted access. If you are using local network storage, this is the responsibility of your network administrators. If you are using a third-party cloud storage solution, then it’s the responsibility of your service providers, and you should check their documentation to ensure their security meets your needs.

## Organising a DDS

One of the problems with a DDS is how to keep it well organised. We all know how difficult it is to organise our own work and our own files – most of us at some point will have overflowing mailboxes, full in-trays and will have spent a good half hour or so searching for that file that we know is “somewhere”! Imagine how much more difficult it is to keep a shared resource organised, especially when several researchers are adding and editing files.

## Folder Structure

It is likely that you will have a folder structure for storing your files. Some systems use tags instead of folders, allowing you to assign multiple tags to each file. While tags are a much more flexible system, folders are more common and generally easier to manage. This is likely because it’s much easier for us to imagine a file as a physical document, placed in a single place inside a folder system.

It’s important to clearly structure your DDS, and this means having clear names for your folders. We recommend including a document at the root of your DDS that describes the folder structure, so everyone in the team is aware of what should go inside each folder.



Some people favour deeply nested folders, while others prefer shallow folder structures. Ultimately, it doesn't really matter how you organise your files, as long as you can effectively communicate the structure to your team and everyone agrees on where files should go.

## Naming Conventions for Files in the DDS

We suggest you develop a naming convention for files within the DDS. One method that is often used is to include the date within the file name. For example, Figure 1 shows a list of files where the name always starts with the date. The date here is in the form YYYY-MM-DD. We recommend this format to make it easy to sort – here, sorting by filename puts the files in date order. You can see that there are two documents called “Dataverse.docx” but one includes the date as 2017-12-07 and the other has the date as 2017-12-11. Don't rely on the modification date as this often picks up the date a file was moved or copied; if you have any Access databases the date modified will change whenever the database was opened regardless of whether or not any changes were made.

Figure 1 - Using Date in the file name

Name	Date modified	Type	Size
2017-12-07 Dataverse.docx	07/12/2017 09:03	Microsoft Word D...	13 KB
2017-12-07 DDS.docx	07/12/2017 10:28	Microsoft Word D...	14 KB
2017-12-07 Metadata.docx	07/12/2017 10:49	Microsoft Word D...	13 KB
2017-12-07 Tools for Research Projects.docx	07/12/2017 10:41	Microsoft Word D...	14 KB
2017-12-07 Training Manual Example using ODK.docx	07/12/2017 14:50	Microsoft Word D...	1,500 KB
2017-12-08 Introduction to Dataverse.docx	11/12/2017 09:46	Microsoft Word D...	256 KB
2017-12-11 Data and Document Storage.docx	11/12/2017 10:03	Microsoft Word D...	256 KB
2017-12-11 Dataverse.docx	11/12/2017 10:07	Microsoft Word D...	14 KB
2017-12-11 Introduction to Dataverse.docx	11/12/2017 10:04	Microsoft Word D...	256 KB

Such naming conventions may seem either obvious or unnecessary, but not defining them at the start can lead to a mess later on. Figure 2 shows a number of files stored together in the same folder. These files are actually different versions of the same document, but it would be very difficult to find the “correct” version or understand what the differences are between the files.

Figure 2 - Versions of the same document

pr-fig1.u.shg	01/02/2000 16:14	SHG File
p-tgs-2000-10.doc	25/10/2000 17:50	Microsoft Office Word 97 - 2003
p-tgs-2000-05.doc	18/05/2000 17:00	Microsoft Office Word 97 - 2003
p-tgs-99.doc	18/10/1999 13:34	Microsoft Office Word 97 - 2003
p-tgs-00.doc	23/03/2000 11:45	Microsoft Office Word 97 - 2003
p-tgs.doc	16/05/2001 11:59	Microsoft Office Word 97 - 2003
PresentingResults.doc	21/01/2000 12:40	Microsoft Office Word 97 - 2003
presenting results- revised by RC.doc	15/11/1999 08:54	Microsoft Office Word 97 - 2003
Presentation.doc	04/04/2000 15:07	Microsoft Office Word 97 - 2003
Presentation booklet.doc	11/01/2000 12:14	Microsoft Office Word 97 - 2003
present.doc	03/02/1998 13:38	Microsoft Office Word 97 - 2003
h-tgs.doc	16/05/2001 12:02	Microsoft Office Word 97 - 2003
h-gfp.doc	15/04/1998 08:47	Microsoft Office Word 97 - 2003
pr-hpura.tif	12/12/2002 18:02	TIF File

## Sorting out the mess

Even with an agreed naming convention, situations like in figure 2 are common. If you manage to catch a mess before it gets too big, you might be able to organise the files by talking to your team and figuring out what the files should be called. But that's not always possible, and a big mess might need a lot of time to sort out.

One suggestion is to create a 'backlog'. Move the mess into a separate space dedicated to your backlog. Then, if you haven't already got a system in place, set up your DDS as if you were starting from scratch. Any new documents and data can go into the new clean structure with names following the agreed convention. Doing this means that your backlog doesn't hold up new work, and you can spend a bit of time each day or week sorting it out. Ideally, you will eventually get through your backlog and have it all fully organised into your main system.

## Data Storage Models

Many data storage models can be used for the DDS. A decentralised system where every person on the team keeps files but each person deposits files into a single place for safekeeping and as a depository of the most up to date version of all data and documents could be used. This has the advantage of freedom but the major disadvantage that ensuring completeness and up-to-date documents is very difficult.

With current technology a centralised system is quite appealing because it makes it easy to ensure completeness and up-to-date documents without much need for coordination of people. One disadvantage to a central store to which everyone has access is that it can easily become a dump which is then very difficult and time-consuming to sort out. To avoid a "dump" you might consider appointing someone as the "custodian" of the DDS. Data and documents go to and from the DDS via this person. An alternative would be to give everyone read access but only give write access to the custodian. A compromise solution is to have a shared development folder where everyone has read/write access but also a folder which has read-only access where final versions of files are stored.

## Ownership of the DDS

There is a distinction between the ownership of the DDS and the ownership of the data and documents stored within. A DDS system might be stored on OneDrive, Dropbox, Google Drive or an internal system from your organisation. This does not mean that all the data stored within it belong to those institutions.

Many researchers still have issues over sharing "their" data, somehow believing that in doing so they are giving away their rights and someone else will get the credit for their work. However, you should bear in mind that ownership is generally defined by a contract between the funding agency and the researcher or organisation. Check your contract to ensure you know what this is. Also, it is a good idea to draw up agreements that can be signed by all members of the project team. This ensures that everyone knows where they stand with respect to data ownership. Remember that the Intellectual Property Rights (IPR) of the contents of the DDS remain the same regardless of where the DDS is stored and who has access to it. Ownership does not change simply because you have placed the files into a shared location.

Ultimately, a shared DDS requires some degree of trust. Remember, though, this works both ways: if you are not willing to share your data and documentation with others, you cannot expect them to share their files with you.

## What System(s) Should you use?

Where you store your DDS depends on the resources and local skills you have available.

### A Cloud Storage Service

If you have good internet access, an easy solution is to use a third-party cloud storage solution. As of 2018, the three main contenders are:

- OneDrive / SharePoint
- Dropbox
- Google Drive.

Each company offers a variety of plans, including ones designed for business use. There are a host of other options too and each service offers similar features.

### Benefits

Benefits of a cloud storage system include:

- You can access your DDS anywhere you have an internet connection, making it useful for remote teams;
- Responsibility for server maintenance is passed to the service provider;
- A good service provider has the benefit of scale – they manage far more storage space than your project, so can provide a much cheaper and more stable system than a custom-purpose server.
- Your master files are not stored on any one person's computer, so you are protected against hardware failures.

### Issues

A common concern about cloud storage options is that of security and privacy. Using any internet-based system does increase your security risks slightly, simply by the fact of your data being accessible from more locations. All good systems offer fully encrypted data transfers and encrypted storage. For example, Dropbox uses SSL / TLS to encrypt all data in transit and encrypts data on its servers with 256-bit AES – an industry recognised standard. We recommend carefully reading the documentation for the service you are considering to ensure it meets the requirements of your project.

Organisations may have policies indicating what can and cannot be placed onto a third-party system. For example, a university might require that any personal data about human subjects not be placed onto a third-party system to comply with data protection laws.

### A Shared Network Drive

If your team are in the same location, or can access a local network via a VPN, you could use a shared network drive to which all team members have access.

This is a good solution if you are required to keep data and documents “on-premises”, or do not have a stable internet connection. It means that your team (or someone in your organisation) has responsibility to manage the infrastructure.

With the current pace of change in technological development new solutions are appearing all the time. The important thing is not which technology to use but rather to be aware that establishing a DDS is highly beneficial for a research project and making the managerial decisions that will make use of the best technology available to help achieving good management of the data and documents of a research project.

## Summary

The Data and Document Store is a system to help you keep all your project files together in a centralised location. A well-organised DDS means that team members can always access the latest documents and data and data integrity is preserved. Archiving at the end of the project is made quicker and easier.

Remember though, there is no special software involved and there is certainly no magic wand to organise your files. As a team you must decide on the structure of your DDS to ensure it becomes a useful resource and not just a file dump.

## Associate Videos

Videos accompanying the original release of the CCAFS Data Management Support Pack in 2013 are available as a playlist on the Statistical Services Centre YouTube Channel at <https://www.youtube.com/channel/UCs7EU95YMihvNozJKCD92xQ/playlists>. These videos have not been updated since the original release but are mostly still relevant.

In particular the playlist includes the following videos associated with Data and Document Storage:

### **Introduction to Data and Document Storage -**

[https://www.youtube.com/watch?v=4CQtJbg\\_Qms&index=6&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpi](https://www.youtube.com/watch?v=4CQtJbg_Qms&index=6&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpi)

### **Ownership Issues with Data and Document Stores -**

<https://www.youtube.com/watch?v=ML3UXLzsqRw&index=8&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpi>

### **Data and Document Store Organisation -**

[https://www.youtube.com/watch?v=MMagU\\_77rdI&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpi&index=7](https://www.youtube.com/watch?v=MMagU_77rdI&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpi&index=7)

## Reference

### 18. Introduction to Dropbox

- *Data Management*
  - *Data Storage*
- *Using the Data*
  - *Sharing and Access internally*
- *Reference*
- *When*
  - *Management of research processes*
- *PI, Researcher, Technician*

Please find below the document

# Introduction to Dropbox

## What is Dropbox?

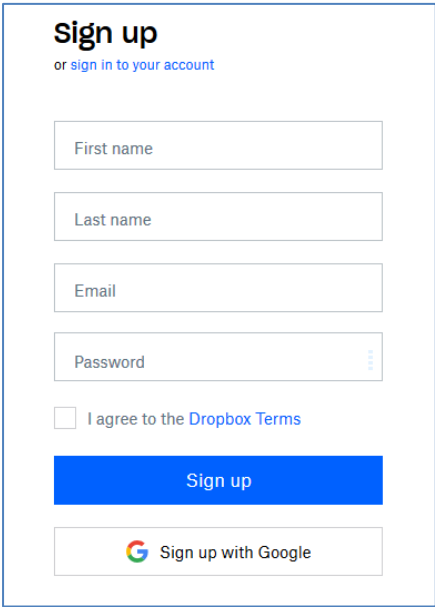
Dropbox is a web-based file hosting service that enables you to:

- Access your files from any computer with an Internet connection; and
- Easily share files with others in your team, even if they are on the other side of the world.

Note the images in this document were captured when using Dropbox for Business.

## Getting Started

To start using Dropbox you first need to create an account for yourself. Go to <https://www.dropbox.com/> and enter your first name, last name, email address in the form on the screen

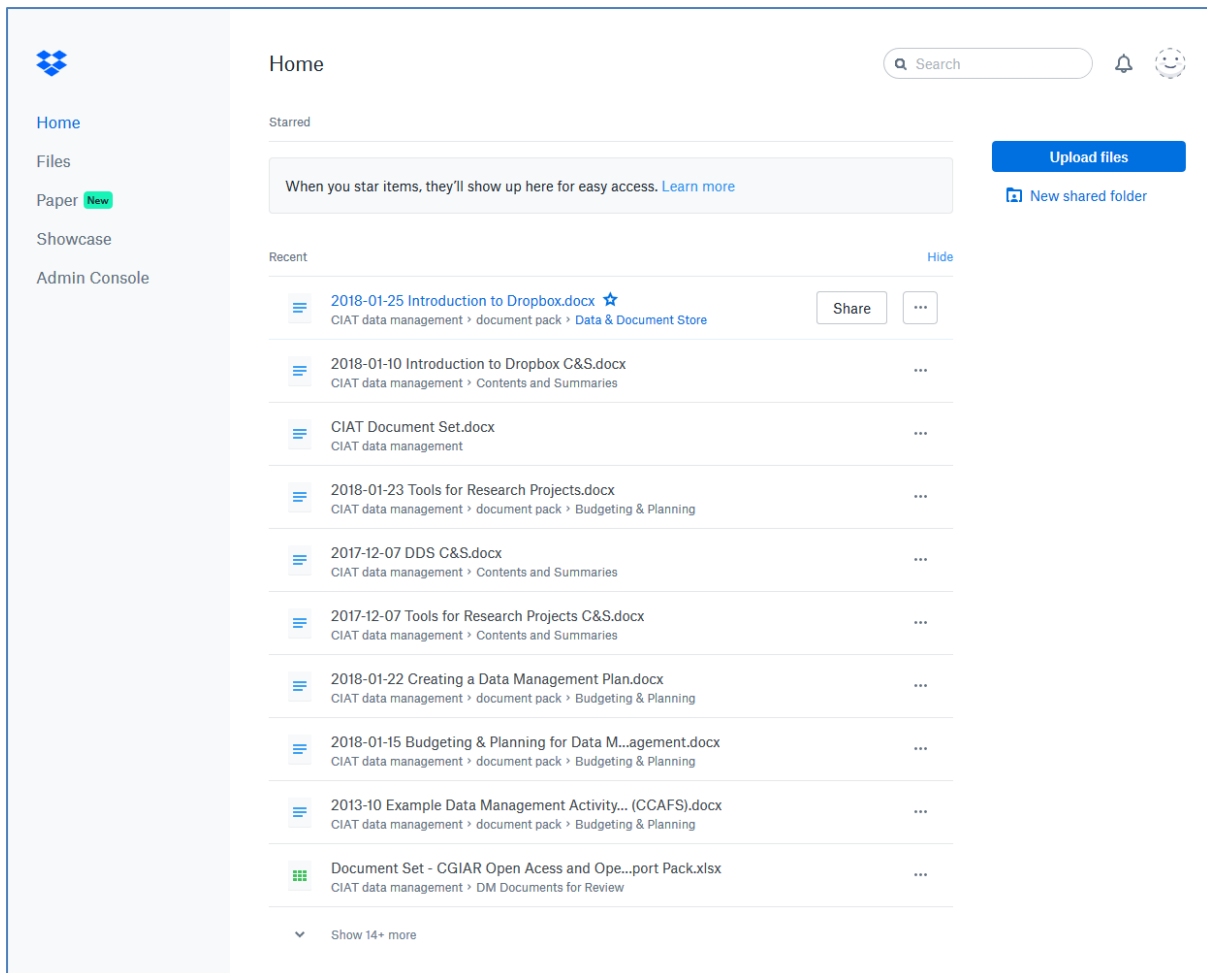


The image shows a screenshot of the Dropbox sign-up form. At the top, it says "Sign up" in bold, with a link "or sign in to your account" below it. The form contains four input fields: "First name", "Last name", "Email", and "Password". Below the "Password" field is a checkbox labeled "I agree to the Dropbox Terms". At the bottom of the form is a blue "Sign up" button and a "Sign up with Google" button with the Google logo.

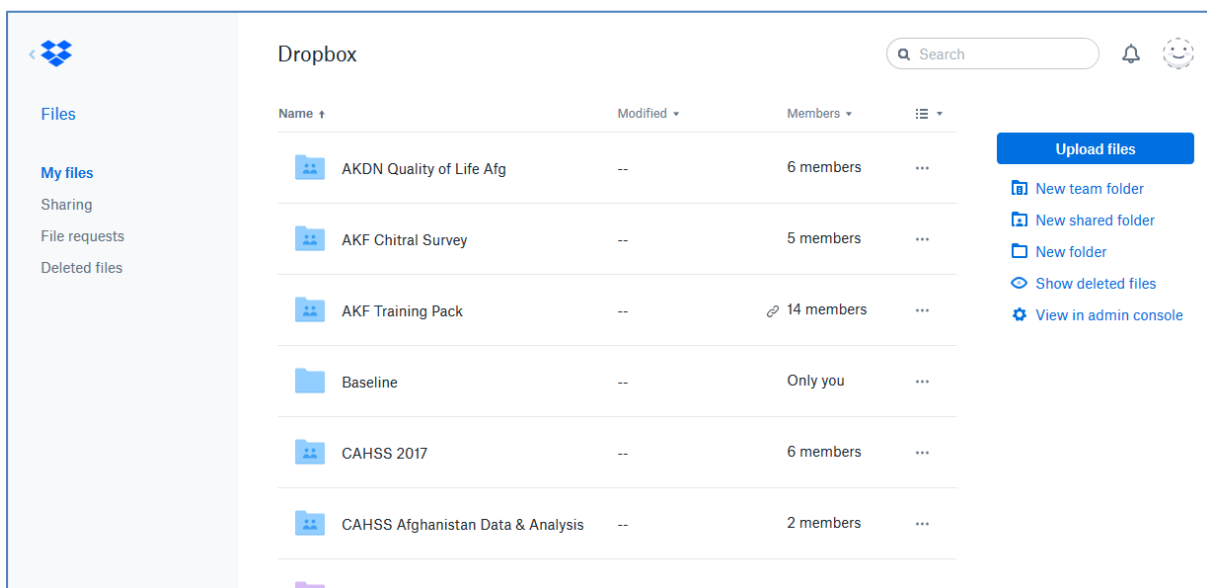
You'll also need a suitable password – make sure you remember your password! You will also need to agree to the Dropbox Terms and Conditions. Then just click on **Sign up** and you're ready to go.

## Working through your browser

Once you have signed in, you can manage your files and folders from your web browser. As shown below you'll initially see a list of files that you have accessed recently. Dropbox allows you to mark items or "star" items so that they will always appear on your home screen when you log in to Dropbox.



Clicking on **Files** in the left-hand column of the screen will take you to your full list of files and folders, this includes folders that you have created yourself and folders that others have shared with you. The image below shows an example.

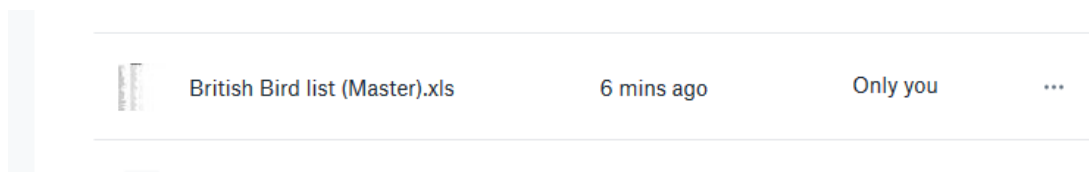


For any files in the list, the date the file was last modified will be shown. The **Members** column shows the number of individuals the folder is shared with – for example the folder “AKF Chitral Survey” is shared with four other people whereas the folder called “Baseline” is not shared at all so is only

available to the current user. You will also notice the symbol next to the 14 members of the AKF Training Pack folder; this indicates that a read-only link has been created for this folder so the files within it can be viewed but not edited by anyone with the link.

## Uploading Files

To update a file, click on **Upload files** in the right-hand column. This will take you to the standard Windows Open File dialog from where you can select a file to upload. The file will be uploaded to the current folder. If you later want to move it to a different folder then click the ellipsis (3 dots) at the end of the row (see below) and select **Move** from the pop-up menu. Select the destination folder for the file from the list of folders you have available.

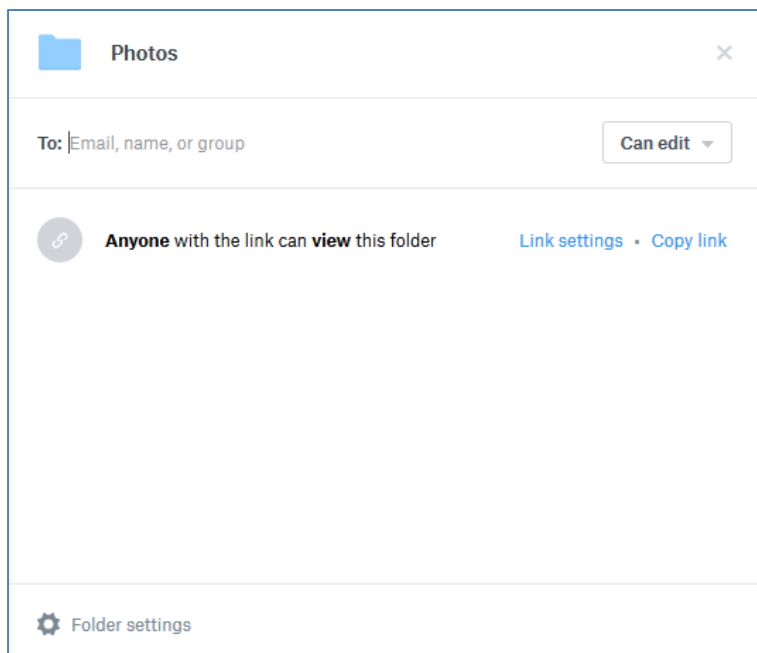


An alternative method for uploading files is simply to drag and drop from Windows Explorer, effectively treating your Dropbox folder as another folder on your hard drive. Remember to hold down the <Ctrl> key if you want to copy rather than move the file.

## Sharing Folders

To share a folder, hover over the folder you want to share and click on **Share** – the Share button will only appear when you hover over the folder. A dialog similar to the one shown below will appear. Here you should enter the email addresses of individuals you wish to share the folder with. If you have Dropbox for Business you can also choose whether you want individuals to be able to edit the folder and files or whether you just want them to be able to view and download the files. The default is to allow others to edit the folder. This includes giving them permission to add and delete files in the folder. For standard Dropbox, the option to allow others to just view the files is not available and you would need to share a link to the file.

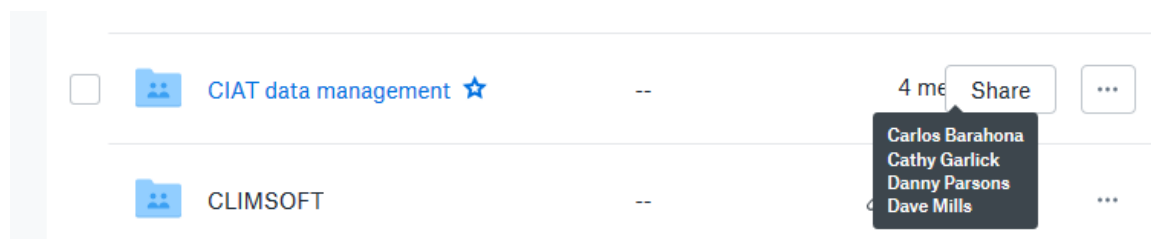




Once you have added at least one email address, a **Share** button will appear at the bottom of the dialog box. There will also be space for you to enter a message should you wish to do so. Each person will receive an invitation email with a link to follow.

### Checking and Changing the Member list

As we saw earlier, when a folder is shared the list will indicate the number of individuals who are currently sharing it. Dropbox refers to these individuals as “members”. To see the current list of members hover over the number of members in the list as shown below.



To change permissions for members or to remove a member from the list, click on Share. This gives you a dialog similar to the one shown earlier but this time you will see the current list of members. Next to each member there is a “Can edit” drop-down and by using this drop-down you can change the permissions for an individual member or remove a member. Remember the option for **Can View** is only available in Dropbox for Business.

**CIAT data management**

4 members

✕

To: Email, name, or group

Can edit ▾

🔗

No link created yet

[Create a link](#)

CB

**Carlos Barahona**

c.e.barahona@stats4sd.org

Can edit ▾

CG

**Cathy Garlick**

c.garlick383@btinternet.co

DP

**Danny Parsons**

danny@stats4sd.org

DM

**Dave Mills**

d.e.mills@stats4sd.org

✓

**Can edit**

People can edit, delete, comment, and add the file to their Dropbox

**Can view**

People can view, download, and comment








Remove

⚙️

Folder settings

## Events

When you share a folder, those you share it with will have full access; i.e. they can read the files, modify them and even delete them. Dropbox keeps a track of events in your folders such as who edited, added or deleted which file in which folder and when. See the example below.

Events		Q S
Event	Time	
 Dave Mills edited the file <a href="#">FIPS App Blog Post v1_Ti...</a> in SSC	2 hrs ago	
 Dave Mills deleted <a href="#">alex riba.png and 18 more files</a> in SSC	3 hrs ago	
 Dave Mills deleted <a href="#">Stationery and 6 more folders</a> in SSC	3 hrs ago	
 Andrew Pinney added the file <a href="#">171214 DFID ECHO ...</a> in DFID Somalia Project Documents	4 hrs ago	
 Carlos Barahona edited the file <a href="#">Staff meeting 2018...</a> in SSC	Yesterday 5:33 PM	
 You edited the file <a href="#">2018-01-23 Tools for Research.....</a> in CIAT data management	Yesterday 4:36 PM	
 Andrew Pinney edited the file <a href="#">Cash_Call_Centre_M...</a> in DFID Somalia Project Documents	Yesterday 2:11 PM	

To display this list of events you will need to change the URL to <https://www.dropbox.com/events>  
Note you will not be able to access this page unless you are already logged in.

## Linking

If you want to be able to share files with others without them being able to make changes, you can create links. In your Dropbox list, items that are linked will be marked with a chain link symbol as we saw earlier.

To create a link to a file or folder open the sharing dialog as though you were going to share the item. Click on **Create a link**. The link is created, and you can then copy the link as shown in the image below. The link will be copied to the clipboard and from there you can email it to individuals. You can share links to users even if they don't use Dropbox themselves.

**CIAT Document Set.docx** 4 members

To: Email, name, or group Can view

Anyone with the link can view this file Link settings • Copy link

- CG** Cathy Garlick - Viewed 1 hr ago  
c.garlick383@btinternet.com Can edit
- DM** Dave Mills - Viewed 2 days ago  
d.e.mills@stats4sd.org Owner
- CB** Carlos Barahona  
c.e.barahona@stats4sd.org Can edit ▾
- DP** Danny Parsons  
danny@stats4sd.org Can edit ▾

File settings

## Installing Dropbox

All the facilities we have looked at so far can be done through the Dropbox website. If you install Dropbox, a folder will appear on your computer which will automatically sync with the web version of your Dropbox.

	Name	Date modified	Type	Size
<ul style="list-style-type: none"> <li>★ Favorites <ul style="list-style-type: none"> <li>Desktop</li> <li>Downloads</li> <li>Recent Places</li> <li>OneDrive</li> <li>Dropbox (SSD)</li> </ul> </li> <li>Libraries <ul style="list-style-type: none"> <li>Documents</li> <li>Music</li> <li>Pictures</li> <li>Videos</li> </ul> </li> <li>Computer <ul style="list-style-type: none"> <li>Windows (C:)</li> </ul> </li> </ul>	AKDN Quality of Life Afg	08/01/2018 16:42	File folder	
	AKF Chitral Survey	08/01/2018 16:42	File folder	
	AKF Training Pack	08/01/2018 16:43	File folder	
	Baseline	08/01/2018 16:37	File folder	
	CAHSS 2017	08/01/2018 16:43	File folder	
	CAHSS Afghanistan Data & Analysis	08/01/2018 16:41	File folder	
	Camera Uploads	09/01/2018 09:59	File folder	
	Cath-David	08/01/2018 16:37	File folder	
	Cathy	25/01/2018 13:55	File folder	
	CCAFS	08/01/2018 16:37	File folder	
	CCAFS - GPS Photos	08/01/2018 16:37	File folder	
	CCAFS BL MethReview	08/01/2018 16:53	File folder	
	CCRP RM IMEP	08/01/2018 16:39	File folder	
	CIAT data management	22/01/2018 16:26	File folder	

You can use this folder like any other folder on your computer. The only difference being that this folder will be available to you from wherever you are, provided you have an Internet connection. To install Dropbox go to <https://www.dropbox.com/> and click on **Download** to download and run the installer.

## Pricing

A Basic Dropbox account is free and allows you up to 2Gb of storage space, but you can gain extra free space by referring your friends. In this way you can boost your storage space to a maximum of 16Gb.

If you need additional space, then Dropbox Plus allows up to 1Tb of space and also provides offline file access and priority email support. Dropbox Professional also offers 1Tb of storage space but has additional features such as being able to share your work with customised branding, visual previews and informative captions.

As well as these individual plans there are several team plans available as Dropbox Business.

See the website for further information and for up-to-date prices.

## Summary

So, to summarise, Dropbox is an easy to use and very popular file hosting service that gives you access to your files through the web. It enables easy sharing of files both as full access shares or read-only links.

## Associate Videos

Videos accompanying the original release of the CCAFS Data Management Support Pack in 2013 are available as a playlist on the Statistical Services Centre YouTube Channel at <https://www.youtube.com/channel/UCs7EU95YMjhvNozJKCD92xQ/playlists>. These videos have not been updated since the original release but are mostly still relevant.

In particular the playlist includes the following video on Introduction to Dropbox - <https://www.youtube.com/watch?v=kvMkh4sIKCU&index=9&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpj>

# Fieldwork

## Main

### 19. Documents needed for Survey Fieldwork and Training

- *Communications*
  - *Support & Guidance*
- *Data Management*
  - *Data Collection*
- *Using the Data*
  - *Understanding of the DM Process*
  - *User manuals/documentation*
- *Main*
- *When*
  - *Decisions while designing*
  - *Management of research processes*  
*PI, Researcher, Technician*

Please find below the document

# Documents for Survey Fieldwork and Training

## Introduction

This document is intended to provide information on the types of documents that are needed for survey fieldwork and training of field staff. This includes detailed manuals. You are likely to need a general field manual which can be used both for training purposes and as a reference manual for staff while they are in the field. Depending on the number of extra tasks the supervisors are expected to carry out, it might be useful to also have a separate Supervisors' manual.

If using paper questionnaires, you will obviously need copies of the questionnaire in the appropriate language(s) as well as a separate code book for coded questions. A comprehensive checklist is also useful, and enumerators and supervisors can use this for carrying out visual consistency checks on the completed questionnaires.

You will need to have consent forms for the respondents and it might also be useful to have some printed information about the project that you can leave with the respondents.

This data management pack includes examples of some of these documents.

## Fieldwork & Training Manual

This manual should include all the information that field teams will need to know. This should include some background information about the project, sampling information, and details about the questionnaire. If you are using digital data collection you will need to explain how to use the mobile devices, how to record and correct the data, plus general information about using the mobile device such as battery conservation etc. If using paper questionnaires, the manual should include details on completing the questionnaire and carrying out visual data checks.

## Supervisor Manual

The supervisor is likely to have additional tasks. For example, for digital data collection you may have decided that only the supervisor can upload data to the server. The supervisor would also be responsible for ensuring the questionnaires have been completed correctly and completely.

For this section of the Data Management Pack we have the following examples:

### *CCAFS Training Manual for Field Supervisors*

This is the manual that was used for the Household Baseline Survey. The manual introduces the study before going on to talk about site selection and sampling. It gives information about each section of the questionnaire detailing the information that is being collected and the reasons for collecting that information.

The manual then lists the roles and responsibilities of field staff and how they should operate while in the field. It discusses the codebook and how that should be used. Finally, it goes through the questionnaire section by section and question by question.

### *Training Manual Example using ODK*

This example training manual was derived from one used on a research project where survey data were collected using ODK. It is divided into a Fieldwork Section and a Supervisor Section though the Supervisor Section could be saved as a separate manual. The manual covers how to use ODK Collect on the mobile devices and discusses the different types of question being used, skip patterns in the questionnaire, etc.

The manual should then go through the questionnaire in detail, question by question, and finish with a list of checks that should be carried out by the enumerator.

The Supervisor Section includes details about preparing the devices for use, installing applications on the devices, adding forms to the device etc. It also covers how to use Bluetooth connections to backup the data; in the research project that this manual was taken from, the supervisor would take a copy of the data from each member of his/her team at the end of each day. Only after checking the data for obvious errors would the supervisor then upload the data to the aggregate server.

## Questionnaire

If you are using paper questionnaires you will obviously need an ample supply of questionnaires in the appropriate language(s). Wherever possible it is best to avoid the situation where enumerators are translating “on-the-fly” during the interview as there is no guarantee the translation will be correct or will convey the correct meaning. In this section of the pack we have included the questionnaire for the CCAFS Household Baseline Survey.

## Code Book

In many cases codes used in the questionnaire will appear in the questionnaire itself. However, when there are many variables that use the same set of codes, or there is a particularly long set of codes for some questions, then you may need a separate code book. For the CCAFS Household Baseline Survey the questionnaire included many of the codes but codes for ethnicity, crops and livestock were listed in a separate code book. Crop and livestock codes were used for many variables and the ethnicity codes varied according to the site. This code book is included in this pack.

## Checklist

A list of consistency checks may be very useful for field staff to refer to when they are carrying out visual checks of the data after the interview. This might include for example:

- Check that parents are at least 15 years older than their children;
- Check that planting date comes before harvest date;
- Check that those with a university education must be at least 18 years of age;
- Etc.

The same list can be used when designing the data entry systems and when carrying out quality control checks on the data prior to analysis.

## Consent Form

For all surveys you will need to give respondents information about the study and give them the opportunity to decide whether or not they want to take part. Informed consent is an important part of carrying out survey interviews. We have included an example consent form that can easily be adapted to suit the needs of your project. We also suggest having some information about the project that you can leave with the respondents.



## 29. CG Core Metadata Schema

- *Data Management*
  - *Archiving*
  - *Maintenance*
- *Communications*
  - *Data dissemination/publicity*
- *Using the Data*
  - *User manuals/documentation*
- *Main*
- *When*
  - *Management of research processes*
  - *Delivery of research products*
- *PI, Researcher, Technician*

Please find below the document

---

# **CG Core Metadata Schema and Application Profile**

---

**Draft version Beta 1  
23 November 2016**

*About this document:*

This document is designed to present and offer guidance for using CG Core, the set of metadata elements used by CGIAR Research Center and CRP repositories, in order to facilitate cross-repository searching and enhance discovery of CGIAR information products through Open Access.

## CG Core Metadata Schema

The following metadata elements make up the “CG Core” metadata schema, intended to be the minimum set of elements applicable across CGIAR Centers, data streams, and formats. Application of the CG Core to Center publication and data repositories and relevant databases will enable consistent annotation of final research products, to enable adherence to OA-OD under the “FAIR” principles: Findability, Accessibility, Interoperability, and Reusability. CG core will also allow meta-searching and indexing across CGIAR repositories and databases and inter-linking across multiple resources.

This schema is closely aligned with Dublin Core, a generic and widely-adopted metadata schema that been in use since the mid-1990s; and to the Data Documentation Initiative (DDI) by the DDI Alliance that is in usage by many CG Centres through Dataverse. The generic nature of Dublin Core and DDI make them ideally-suited for adoption for a wide variety of purposes, yet it also requires customization in order to offer meaningful, consistent description of materials within a particular context, sector, or subject area and for a particular type of information product.

CG Core is designed to follow Dublin Core and DDI as much as possible, with additional elements or attributes incorporated to capture and share CGIAR-specific administrative information as well as descriptive details that are integral to agriculture and food policy research.

### Element Status: “Required,” “Required When Applicable,” and “Optional”

Ideally, all of the elements for CG Core will be included in each CGIAR repository in order to promote alignment across repositories. However, it is not always possible or applicable to include content in each element. Thus, “**required**” elements should be populated with applicable content in each record of a repository.

“**Required when applicable**” is used for elements that might not always have metadata in each record – such qualifiers might not always be applicable for all records.

Centers/CRPs are encouraged to use the elements that are listed as “**optional**,” but metadata should only be incorporated into this field as it is appropriate and based on Center/CRP discretion.

### Using the CG Core Element Set

Guidance for the use of the metadata elements included in the CG Core metadata schema is included below. The business rules, guidelines for usage, and examples are included for each element.

“Schema links” indicates how the elements maps to Dublin Core and DDI, or if it is only a CG Core element. Text in “*Courier New*” represents the element names as they would appear to end users or examples of the metadata that would be found in each field. All other information is intended as guidance for repository, data, knowledge, and metadata managers.

Several elements are designed to use common vocabularies and lists. Ideally, these terms will be incorporated into repositories as “pick lists” or as autocomplete/suggested text. “CGIAR Lists” is indicated for those lists unique to CGIAR. Terms for the “Subject” element should come from more widely used vocabularies like AGROCOV, CABI Thesaurus or the Global Agricultural Concept Scheme.

## CG Core Elements

<b>Element:</b>	<b>Title</b>
<b>Status:</b>	Required
<b>Tag:</b>	cg.title (Multiple element)
<b>Schema links:</b>	Dublin Core (dc.title), DDI (codebook.docDscr.citation.titlStmt.titl)
<b>Description:</b>	Full official or unofficial title of the information product (document, data set, image, etc.) Can be used as a repeating field to include alternate titles. Used for all types of research outputs and repository materials.
<b>Format:</b>	Follow standard title formatting for capitalization and punctuation. For articles, follow formatting used by the article's publisher.
<b>Examples:</b>	"Managing for timber and biodiversity in the Congo basin" "A 2007 Social Accounting Matrix for China" "2012 Global Hunger Index Data"

<b>Element:</b>	<b>Creator</b>
<b>Status:</b>	Required
<b>Tag:</b>	cg.creator (Multiple element)
<b>Schema links:</b>	Dublin Core (dc.creator), DDI (codebook.docDscr.citation. rspStmt. AuthEnty)
<b>Description:</b>	The author(s), researcher(s), scientist(s) responsible for producing the information product. Indicate the Center where the Center is the corporate author. When the creator is a person indicate use the attribute "Affiliation" to indicate the affiliation.
<b>Format:</b>	Use Center-specified format. Recommended formats include: Last Name, First Name Last Name, First Initial Last Name, First Initial Middle Initial List primary author(s) first – use same order as listed on the publication/research product.
<b>Examples:</b>	"Nasi, Robert" "Smith, B" "International Center for Tropical Agriculture (CIAT)"

<b>Element:</b>	<b>Creator</b>
<b>Attribute:</b>	<b>ID</b>
<b>Status:</b>	Required when applicable – should be used as appropriate when a Center has implemented ORCID or other type of author identifier.
<b>Schema links:</b>	DDI (codebook.docDscr.citation. rspStmt. AuthEnty [Attribute ID])
<b>Description:</b>	Used if ORCID, SCOPUS, or other type of creator ID scheme is in use
<b>Format:</b>	Use format as specified by the source of the identifier with an @ to indicate the source.
<b>Examples:</b>	"Type=0000-0003-3347-861X@ORCID" "Type=0000-0002-7628-3348@SCOPUS"

<b>Element:</b>	<b>Creator</b>
<b>Attribute:</b>	<b>Affiliation</b>
<b>Status:</b>	Required when applicable
<b>Schema links:</b>	DDI (codebook.docDscr.citation. rspStmt. AuthEnty [Attribute affiliation])
<b>Description:</b>	This is the affiliation of the creator. Can accommodate various separated by coma
<b>Format:</b>	Use format as specified by the source of the identifier.
<b>Additional Details:</b>	None
<b>Examples:</b>	"Affiliation=Wageningen University"

<b>Element:</b>	<b>Subject</b>
Status:	Required
Tag:	cg.subject (Multiple element)
Schema links:	Dublin Core (dc.subject), DDI (codebook.docDscr.subject.keyword)
Description:	The subject matter of the research, technologies tested, crops involved in the research, methodologies, etc.
Format:	Single words or short phrases. Use controlled vocabularies (see attribute vocab)
Additional Details:	Further work is needed around harmonization of Center-specific terms where terms overlaps. Further work also needed to harmonize CG subjects with AGROVOC when possible.
Examples:	Cattle, Dairy, Maize

<b>Element:</b>	<b>Subject</b>
Attribute:	<b>vocab</b>
Status:	Optional
Schema links:	Dublin Core (dc.subject [Attribute xsi:type]), DDI (codebook.docDscr.subject.keyword [Attribute vocab])
Description:	Vocabulary used for each subject term.
Format:	Code or name of the vocabulary: AGROVOC (AGROVOC Multilingual agricultural thesaurus): <a href="http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus">http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus</a> CAB (CABI Thesaurus): <a href="http://www.cabi.org/cabthesaurus/mtwdk.exe?yi=home">http://www.cabi.org/cabthesaurus/mtwdk.exe?yi=home</a> GACS (Global Agricultural Concept Scheme): <a href="http://browser.agrisemantics.org/gacs/en/">http://browser.agrisemantics.org/gacs/en/</a>
Examples:	"vocab=AGROVOC" "vocab=CAB" "vocab=GACS"

<b>Element:</b>	<b>Description</b>
Status:	Optional, see note below regarding discoverability.
Tag:	cg.description (Single element)
Schema links:	Dublin Core (dc.description) DDI (codebook.stdyDscr.stdyInfo.abstract)
Description:	Abstract, short or long description of information/data product. Especially important for datasets, software, journal articles, working papers, reports, and other types of written materials.  Can be in a language other than the original language in which an item was produced.
Format:	Short description, a few sentences, or longer paragraph-style text.
Additional Details:	Descriptive details significantly improve discoverability via search engines such as Google and Bing, and will aid interlinkages between related resources at the meta-search/indexer level.  Descriptions can be provided in multiple languages if appropriate and available.
Example:	"Drought is one of the major constraints affecting food security and livelihoods of more than two billion people that reside on dry areas which constitute 41% of the world's land surface. Drought is defined as deficiency of precipitation over an extended period of time resulting in water scarcity. Our best minds should be concentrated where the greatest challenges lie today - on discoveries and new solutions to cope with the challenges facing dry areas particularly drought and water scarcity. In addition to facing severe natural resource constraints caused by the lack of water in many of the developing world's drylands, we also have to cope with rapid growth of the younger segment of the growing population, and high levels of poverty. Coping with drought and water scarcity are critical to address major development challenges in dry areas namely poverty, hunger, environmental degradation and social conflict. Drought is a climatic

event that cannot be prevented, but interventions and preparedness to drought can help to: (i) be better prepared to cope with drought; (ii) develop more resilient ecosystems (iii) improve resilience to recover from drought; and (iv) mitigate the impacts of droughts. Preparedness strategies to drought include: (a) geographical shifts of agricultural systems; (b) climate-proofing rainfall-based systems; (c) making irrigated systems more efficient; (d) expanding the intermediate rainfed-irrigated systems. The paper presents successful research results and case studies applying some innovative techniques where clear impact is demonstrated to cope with drought and contribute to food security in dry areas.”

<b>Element:</b>	<b>Publisher</b>
<b>Status:</b>	Required when applicable – i.e. for peer-reviewed journal articles (including data articles)
<b>Tag:</b>	cg.publisher (Single element)
<b>Schema links:</b>	Dublin Core (dc.publisher), DDI (codebook.docDscr.citation.prodStmt.producer)
<b>Description:</b>	Entity responsible for publication, distribution, or imprint – not the journal title, but the publisher
<b>Format:</b>	Use standard capitalization
<b>Examples:</b>	“Academic Journals” “Elsevier” “PLOS”

<b>Element:</b>	<b>Contributor</b>
<b>Status:</b>	Required
<b>Tag:</b>	cg.contributor (Multiple elements)
<b>Schema links:</b>	Dublin Core (dc.contributor), DDI (codebook.docDscr.citation.rspStmt.othId)
<b>Description:</b>	Person, organization, or service making contributions to the information product.
<b>Additional Details:</b>	<p>Use the attribute “Type” to specify the type of contributor: Person, Organization, Center, CRP, Partner, Funder, Project, or Project Lead Institution .</p> <p>In the case of a person use the attribute “Role” to indicate the role in the production of the information product.</p> <p>Use the attribute “Affiliation” to indicate the affiliation of the person</p>
<b>Format:</b>	<p>Free entry text except for types Centre and CRP when is a fixed list of entries.</p> <p>When the type is “Centre” the following “CGIAR list” apply:</p> <p>“AfricaRice” “Bioversity International” “Center for International Forestry Research (CIFOR)” “International Center for Agricultural Research in the Dry Areas (ICARDA)” “International Center for Tropical Agriculture (CIAT)” “International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)” “International Food Policy Research Institute (IFPRI)” “International Institute of Tropical Agriculture (IITA)” “International Livestock Research Institute (ILRI)” “International Maize and Wheat Improvement Center (CIMMYT)” “International Potato Center (CIP)” “International Rice Research Institute (IRRI)” “International Water Management Institute (IWMI)” “World Agroforestry Centre (ICRAF)” “WorldFish”</p> <p>When the type is “CRP” the following “CGIAR lists” apply:</p> <p>“CGIAR Research Program on Agriculture for Nutrition and Health (A4NH)” “CGIAR Research Program on Aquatic Agricultural Systems (AAS)”</p>

"CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS)"  
 "CGIAR Research Program on Dryland Cereals (Dryland Cereals)"  
 "CGIAR Research Program on Dryland Systems (Dryland Systems)"  
 "CGIAR Research Program on Forests, Trees and Agroforestry (ForestsTreesAgroforestry)"  
 "CGIAR Research Program for Managing and Sustaining Crop Collections (Genebanks)"  
 "CGIAR Research Program on Grain Legumes (Grain Legumes)"  
 "CGIAR Research Program on Rice (GRiSP)"  
 "CGIAR Research Program on Integrated Systems for the Humid Tropics (Humidtropics)"  
 "CGIAR Research Program on Livestock and Fish (Livestock and Fish)"  
 "CGIAR Research Program on Maize (MAIZE)"

	"CGIAR Research Program on Policies, Institutions and Markets (Policies, Institutions and Markets)" "CGIAR Research Program on Roots, Tubers and Bananas (RTB)" "CGIAR Research Program on Wheat (WHEAT)" "CGIAR Research Program on Water, Land and Ecosystems (WLE)"
<b>Examples:</b>	CGIAR International Water Management Institute (IWMI) CGIAR Research Program on Water, Land and Ecosystems (WLE) University of California, Davis Erick Rutto

<b>Element:</b>	<b>Contributor</b>
<b>Attribute:</b>	<b>Type</b>
Status:	Required
Schema links:	DDI (codebook.docDscr.citation.rspStmt.othId [Attribute "type"])
Description:	Type of contributor
Format:	Fixed list of entries: Person Organization Center CRP Partner Donor Project, or Project Lead Institution
Additional Details:	List of entries could come from a controlled list.
Example:	"type= Person"

<b>Element:</b>	<b>Contributor</b>
<b>Attribute:</b>	<b>Role</b>
Status:	Required when applicable
Schema links:	DDI (codebook.docDscr.citation.rspStmt.othId [Attribute "role"])
Description:	When the type is person the role that person had in the production of the information product.
Format:	Free Text
Examples:	"Role=Editor" "Role=Statistician"

<b>Element:</b>	<b>Contributor</b>
<b>Attribute:</b>	<b>Affiliation</b>
Status:	Optional
Schema links:	DDI (codebook.docDscr.citation.rspStmt.othId [Attribute "affiliation"])
Description:	When the type is person then the affiliation such person has.
Format:	Free Text
Examples:	"Affiliation =Wageningen University" "Affiliation =Kenyan Agricultural and Livestock Research Organization (KALRO)"

<b>Element:</b>	<b>Date</b>
Status:	Required

Tag:	cg.date (Multiple elements)
Schema links:	Dublin Core (dc.date or dc.dcterms:available) DDI (codebook.docDscr.citation.prodStmt.prodDate or codebook.docDscr.citation.distStmt.distDate) depending on attribute "type"
Description:	Hold two different types of date: production date and distribution/availability date. This is managed by the attribute "type". Production date: The date when the information product was created in its final form to be published. Distribution date: In cases when the information product has an embargo this date indicates when it would be available.
Format:	It is recommended that repositories adopt one of the following three formats: YYYY-MM-DD (confirm to ISO 8601) YYYY YYYY-MM
Additional Details:	Other types like pre-print date could be added to the type.
Examples:	"2011-12-01" "2015" "2015-12"

<b>Element:</b>	<b>Date</b>
<b>Attribute:</b>	<b>Type</b>
Status:	Required
Schema links:	When type is "production": Dublin Core (dc.date), DDI (codebook.docDscr.citation.prodStmt.prodDate) When type is "distribution": Dublin Core (dc.dcterms:available) DDI (codebook.docDscr.citation.distStmt.distDate)
Description:	Used to know whether a date is Production or distribution
Format:	List of values: Production Distribution
Additional Details:	For publications, datasets, and other types of information products that are deposited upon completion (Production date) but may not be publicly-accessible immediately upon deposit, the distribution date indicates the date upon which the item will be publicly accessible.
Example:	"type= Distribution"

<b>Element:</b>	<b>Type</b>
Status:	Required
Tag:	cg.type (Single element)
Schema links:	Dublin Core (dc.type), DDI (codeBook.stdyDscr.stdyInfo.sumDscr.dataKind)
Description:	Nature or genre of the item, content, or information product
Format:	Use singular words or phrases ("Image"). Use terms from the list below.
Additional Details:	The list of terms below can be expanded to include other types of research outputs and information products, based on the contents in a particular repository. The list of terms presented here includes materials specifically covered in the CGIAR Open Access and Data



	Management Policy as well as other research outputs and data/information products commonly collected and disseminated via CGIAR repositories.
List of Terms:	<p>"Audio"  "Book"  "Book Chapter"  "Dataset"  "Extension Material"  "Image"  "Map"  "Model"  "Peer-reviewed journal article"  "Policy Brief"  "Report"  "Software"  "Source Code"  "Thesis"  "Training Material"  "Video"</p>

<b>Element:</b>	<b>Format</b>
Status:	Required
Tag:	cg.format (Single element)
Schema links:	Dublin Core (dc.format), DDI (codeBook.fileDscr.fileTxt.fileType)
Description:	File format of item
Format:	Standard MIME-type identifier for file format
Additional Details:	List available in Wikipedia: <a href="http://en.wikipedia.org/wiki/Internet_media_type">http://en.wikipedia.org/wiki/Internet_media_type</a>
Details:	Some repository systems may pull this information directly from the object and will not require manual input for this element. Dataverse will require manual input.
Examples:	<p>"application/pdf"  "image/jpeg"  "application/vnd.ms-excel"  "application/zip"</p>

<b>Element:</b>	<b>Identifier</b>
Status:	Required
Tag:	cg.identifier (multiple elements)
Schema links:	Dublin Core (dc.identifier or dcterms:bibliographicCitation), DDI (codeBook.docDscr.citation.titlStmt.IDN or codeBook.docDscr.citation.biblCit) depending on attribute "type"
Description:	Unambiguous reference to the information product such as DOI, URI or Human-readable, standard bibliographic citation for the information product.
Examples:	<p>If type is "Identifier"  "http://hdl.handle.net/10568/66578"  "http://oar.icrisat.org/id/eprint/8611"  "http://dx.doi.org/10.1007/s10661-014-4155-1"</p> <p>If type is "Citation"  "Gumma, M K and Kajisa, K and Mohammed, I A and Whitbread, A M and Nelson, A and Rala, A and Palanisami, K (2015) <i>Temporal change in land use by irrigation source in Tamil Nadu and management implications</i>. Environmental Monitoring and Assessment, 187 (1). pp. 1-17. ISSN 1573-2959"  "Schoeman, S.J. 2000. A Comparative assessment of Dorper sheep in different production environments and systems. Small Ruminant Research 36: 137 - 146."</p>

Element:	<b>Identifier</b>
Attribute:	<b>Type</b>
Required:	Required
Schema:	When type is "Identifier": Dublin Core (dc.identifier), DDI (codeBook.docDscr.citation.titlStmt.IDN) When type is "Citation": Dublin Core (dc.dcterms:bibliographicCitation) DDI (codeBook.docDscr.citation.biblCit)
Description:	Type of Identifier
Format:	List of elements: Identifier Citation
Example:	"Type=Citation"

Element:	<b>Source</b>
Status:	Required when applicable
Tag:	cg.source
Schema links:	Dublin Core (dc.source), DDI (codeBook.stdyDscr.method.dataColl.sources)
Description:	The original journal or other type of material where an item was originally published. Used for journal articles, data articles, conference proceedings, etc.
Format:	Journal title Journal/conference title; vol., no. (year)
Examples:	"PLoS One" "Science" "World Development" "Journal of Development Economics" "The American Journal of Clinical Nutrition"

Element:	<b>Language</b>
Status:	Optional
Tag:	cg.language
Schema:	Dublin Core (dc.language)
Description:	Language of the item
Format:	ISO 639-1 (alpha-2) or ISO 639-2 (alpha-3)
Additional Details:	Use for human languages only, not computer/software programming languages. (Use "Subject" instead for software languages.)
Examples:	"EN" "ES" "FR"

Element:	<b>Relation</b>
Status:	Optional
Tag:	cg.relation (multiple elements)
Schema links:	Dublin Core (dc.relation), DDI (codeBook.stdyDscr.othrStdyMat)
Description:	A related resource for example a News Paper article, a Blog, another publication, etc.
Examples:	"http://dx.doi.org/doi:10.1018/S1537592710004081"

Element:	<b>Coverage</b>
Status:	Required when applicable
Tag:	cg.coverage (multiple elements)
Schema links:	Dublin Core (dc.coverage), DDI (several items of codeBook.stdyDscr.stdyInfo.sumDscr depending on the attribute "Type")

Description:	<p>Geospatial coordinates, countries, regions, sub-regions, chronological period. The type of coverage will depend on the attribute "type":</p> <ul style="list-style-type: none"> <li>Geospatial (Defines a geographical point in the space)</li> <li>Region</li> <li>Country</li> <li>Administrative Unit</li> <li>Chronological period (Defines a date)</li> </ul> <p>"Geospatial" type has the following extra attributes:</p> <ul style="list-style-type: none"> <li>X</li> <li>Y</li> </ul> <p>"Chronological period" has the following extra attribute</p> <ul style="list-style-type: none"> <li>Event [Start, End or Single]</li> </ul>
Format:	<p>Mixed depending on the type:</p> <p>Geospatial: X and Y coordinate in decimal degrees.</p> <p>Region: Use UN stats (<a href="http://unstats.un.org/unsd/methods/m49/m49regin.htm">http://unstats.un.org/unsd/methods/m49/m49regin.htm</a>)</p> <p>Country: Use country names from ISO 3166 (<a href="https://www.iso.org/obp/ui/#search">https://www.iso.org/obp/ui/#search</a>)</p> <p>Administrative Unit: Free text</p> <p>Chronological period: Free text however if a date is stored then ISO 8601 should be used.</p>
Additional Details:	<p>Combination of types can be used.</p> <p>Note on Country: For instances in which there is a lack of clarity regarding countries, best practice is to not include a country. Likewise, in cases where research has occurred in politically sensitive areas and including country-level information could be problematic, best practice is to not include such details in the record. Additional guidance in this area will be forthcoming.</p> <p>Note on Administrative unit: In order to facilitate discovery, all records, when applicable should include tagging of information at sub-national level (ideally district-level). In practice, that means attaching GAUL district codes providing coordinates, place names etc. that are clearly identifiable within a district. A growing amount of socio-eco and climate/soil data is representative at district level, so district labels have become a common denominator to relate research outputs and datasets.</p> <p>Note on Geolocation: It is possible to store different points; therefore geolocation can store points and polygon.</p>
Examples:	<p>"Eastern Africa "</p> <p>"Kenya"</p> <p>"Makueni"</p> <p>"2014"</p> <p>"2017"</p>

Element:	<b>Coverage</b>
Attribute:	<b>Type</b>
Status:	Required when applicable
Schema links:	<p>Dublin Core (dc. coverage)</p> <p>DDI:</p> <ul style="list-style-type: none"> <li>Region and Administrative Unit maps to codeBook.stdyDscr.stdyInfo.sumDscr.geogCover</li> <li>Country maps to codeBook.stdyDscr.stdyInfo.sumDscr.nation</li> <li>Geospatial maps to elements:</li> </ul>

	<ul style="list-style-type: none"> <li>○ codeBook.stdyDscr.stdyInfo.sumDscr. boundPoly. Polygon.point. gringLat</li> <li>○ codeBook.stdyDscr.stdyInfo.sumDscr. boundPoly. Polygon.point. gringLon</li> </ul> <p>Chronological period maps to codeBook.stdyDscr.stdyInfo.sumDscr. collDate</p>
Description: Holds the type of coverage Format:	
The following types are allowed:	
	<p style="text-align: center;">Geospatial Region Country Administrative Unit Chronological period</p>
Examples:	<p>“type=Geospatial” “type=Region”</p>

Element:	<b>Coverage</b>
Attribute:	<b>X</b>
Status:	Required when defining a geospatial coverage
Schema links:	Dublin Core (dc. coverage) DDI: codeBook.stdyDscr.stdyInfo.sumDscr. boundPoly. Polygon.point. gringLon
Description:	Holds the X coordinate in a geospatial coverage
Format:	Decimal values
Examples:	“x=0.00037373737”

Element:	<b>Coverage</b>
Attribute:	<b>Y</b>
Status:	Required when defining a geospatial coverage
Schema links:	Dublin Core (dc. coverage) DDI: codeBook.stdyDscr.stdyInfo.sumDscr. boundPoly. Polygon.point. gringLat
Description:	Holds the Y coordinate in a geospatial coverage
Format:	Decimal values
Examples:	“y=0.00037373737”

Element:	<b>Coverage</b>
Attribute:	<b>Event</b>
Status:	Required when defining a chronological period coverage
Schema links:	Dublin Core (dc. coverage), DDI (codeBook.stdyDscr.stdyInfo.sumDscr. collDate [attribute event])
Description:	Used to indicate the type of chronological period
Format:	The following types are allowed: Start: Defines the start of a period. End: Defines the end of a period. Single: Defines one single event.
Examples:	<p>“Event=Start” “Event=End” “Event=Single” ”</p>

Element:	<b>Rights</b>
Status:	Required
Tag:	cg.rights (multiple elements)

Schema:	Dublin Core (dc. rights) DDI (codeBook.docDscr.citation.prodStmt.copyright)
Description:	Rights (i.e. terms of use, intellectual property rights, licensing details, and/or permissions statement) identifying level/degree of Open Access
Format:	See list of statements below.
Additional Details:	Taking into account whether self or externally published, identify (i) the applicable standard open license (preferred for machine-readability) OR identify the key rights re access/use AND (ii) permissions if restrictions apply. For assistance contact r.sara@cgiar.com.
Examples:	<p><i>If externally published including via OA journal (as per publisher contract):</i></p> <p>"CC BY 4.0"  "CC BY-NC 4.0"; permissions ([publisher e-mail])  "© [publisher] All rights reserved; self-archive copy only, permissions ([publisher e-mail])"  "© [publisher]; Non-commercial educational use only; permissions ([publisher e-mail])"  "© [publisher]; Non-commercial use only; permissions ([publisher e-mail])"</p> <p><i>If self-published (as per donor requirement, Center/CRP policy, or preference):</i></p> <p>"CC BY 4.0"  "CC BY-NC 4.0; permissions ([Center e-mail])"  "Access (unrestricted); Re-use (unrestricted)"  "Access (unrestricted); Reuse (non-commercial only); Permissions ([Center e-mail])"  "Access (unrestricted); Reuse (non-commercial, no translations); Permissions ([Center e-mail])"</p>

Field:	<b>Contact</b>
Status:	Optional, strongly encouraged for datasets and data repositories
Tag:	cg.contact (multiple elements)
Schema links:	DDI (codeBook.docDscr .citation.distStmt.contact)
Description:	It is recommended to use a department rather than an individual; intended to provide a point of contact for anyone who has questions or needs further guidance about the dataset or information product connected to the record. Email is stored in the attribute email
Format:	Use the default format provided by institution
Examples:	"Department of Ecology, University of Wageningen"

Element:	<b>Contact</b>
Attribute:	<b>Email</b>
Status:	Required when applicable
Schema links:	DDI: codeBook.docDscr .citation.distStmt.contact [attribute email]
Description:	Holds the email of the contact
Format:	email
Examples:	"email=j.vanetten@wur.nl"

## Implementation

The implementation of CG-Core happens in three stages: 1) The collection of the necessary elements for each information product that is going to be published, 2) Establishing standard vocabularies and, 3) The implementation of CG-Core by the Centre's repositories.

### Collecting the necessary elements

Each information product that required publishing needs to fulfill the required elements of CG Core. The below template summarizes the elements. **Red** means required, **Yellow** means optional. This template can also be used as a checklist for information products already stored in repositories. See Annex I for an example.

Element	Value	Attributes								
		ID	Type	Affiliation	Vocab	Role	X	Y	Event	Email
Title	Red	NA	NA	NA	NA	NA	NA	NA	NA	NA
Creator	Red	Yellow	NA	Yellow	NA	NA	NA	NA	NA	NA
Subject	Red	NA	NA	NA	Yellow	NA	NA	NA	NA	NA
Description	Red	NA	NA	NA	NA	NA	NA	NA	NA	NA
Publisher	Red	NA	NA	NA	NA	NA	NA	NA	NA	NA
Contributor (type = person)	Red	NA	Red	Red	NA	Red	NA	NA	NA	NA
Contributor (type = not person)	Red	NA	Red	NA	NA	NA	NA	NA	NA	NA
Date	Red	NA	Red	NA	NA	NA	NA	NA	NA	NA
Type	Red	NA	NA	NA	NA	NA	NA	NA	NA	NA
Format	Red	NA	NA	NA	NA	NA	NA	NA	NA	NA
Identifier	Red	NA	Red	NA	NA	NA	NA	NA	NA	NA
Source	Red	NA	NA	NA	NA	NA	NA	NA	NA	NA
Language	Red	NA	NA	NA	NA	NA	NA	NA	NA	NA
Relation	Red	NA	NA	NA	NA	NA	NA	NA	NA	NA
Coverage (type = Geospatial)	NA	NA	Red	NA	NA	NA	Red	Red	NA	NA
Coverage (type = Period)	Red	NA	Red	NA	NA	NA	NA	NA	Red	NA
Coverage (type = not Period or Geospatial)	Red	NA	Red	NA	NA	NA	NA	NA	NA	NA
Rights	Red	NA	NA	NA	NA	NA	NA	NA	NA	NA
Contact	Red	NA	NA	NA	NA	NA	NA	NA	NA	Red

### Implementing standard vocabularies and lists

Central to the implementation of CG Core is the usage of standard vocabularies and lists. These common terms ensure that an information product can be associated with others even external to CGIAR. CG Core uses one vocabulary and six lists:

**Subject:** This is the only vocabulary in CG Core. It allows an information product to be linked to others by simple terms like "Cattle" or "Maize". Three controlled vocabularies can be used:

- GACS (Global Agricultural Concept Scheme, <http://browser.agrisemantics.org/gacs/en/>): Mergers AGROVOC, CABI Thesaurus and NAL Thesaurus (National Agricultural Library's Agricultural Thesaurus)
- AGROVOC (<http://aims.fao.org/standards/agrovoc> )
- CABI Thesaurus (<http://www.cabi.org/cabthesaurus/mtwdk.exe?yi=home>)

**CRP and Centre contributors:** These two lists allow linking information products to centres and CRPs. Both lists are controlled by CGIAR System Organization.

**Contributor type:** This list establishes the type of the contributor. This list is controlled by CGIAR System Organization.

**Type:** This list allows the association of different information products by their type. This list is controlled

by CGIAR System Organization.

**Format:** This list allows the association of different information products by their format. IANA Media Types (<https://www.iana.org/assignments/media-types/media-types.xhtml>) controlled lists must be used here.

**Language:** This controlled list link information products by their language. ISO 639-1 (alpha-2) or ISO 639-2 (alpha-3) must be used. [http://www.infoterm.info/standardization/iso\\_639\\_1\\_2002.php](http://www.infoterm.info/standardization/iso_639_1_2002.php)

**Region and Country coverage:** These two controlled lists allow linking information products by singular geographical locations. Two lists must be used:

- For regions use the United Nations Statistics Division- Standard Country and Area Codes Classifications (M49, <http://unstats.un.org/unsd/methods/m49/m49regin.htm>)
- For countries use ISO 3166 (<https://www.iso.org/obp/ui/#search>)

Note: Terms that are not present in CGIAR controlled lists can be requested for addition to the CGIAR System Organization.

## Implementation of CG Core in different repositories

### Dataverse repositories

Dataverse uses DDI for storing metadata. Since CG Core has been aligned to DDI (with the exception of the element "Language") it should be straight forward to implement CG Core, however Dataverse users need to check if the following list of elements are present for each dataset:

Element in CG Core	Element in Dataverse	Vocabulary / list required
Title	codebook.docDscr.citation.titlStmt.titl	
Creator	codebook.docDscr.citation.rspStmt.AuthEnty	
Subject	codebook.docDscr.subject.keyword	YES
Description	codebook.stdyDscr.stdyInfo.abstract	
Publisher	codebook.docDscr.citation.prodStmt.producer	
Contributor	codebook.docDscr.citation.rspStmt.othId [Attributes: Type, Role and Affiliation]	YES
Date (Type=Production)	codebook.docDscr.citation.prodStmt.prodDate	
Date (Type= Availability)	codebook.docDscr.citation.distStmt.distDate	
Type	codeBook.stdyDscr.stdyInfo.sumDscr.dataKind	YES
Format	codeBook.fileDscr.fileTxt.fileType	YES
Identifier (Type= Identifier)	codeBook.docDscr.citation.titlStmt.IDN [Attribute agency=DOI]	
Identifier (Type= Citation)	codeBook.docDscr.citation.biblCit	
Source	codeBook.stdyDscr.method.dataColl.sources	
Language	Not in DDI.	
Relation	codeBook.stdyDscr.othrStdyMat	
Coverage (Type= Geospatial)	codeBook.stdyDscr.stdyInfo.sumDscr.boundPoly.Polygon.point.gringLat and codeBook.stdyDscr.stdyInfo.sumDscr.boundPoly.Polygon.point.gringLon	
Coverage (Type= Region)	codeBook.stdyDscr.stdyInfo.sumDscr.geogCover	YES
Coverage (Type= Country)	codeBook.stdyDscr.stdyInfo.sumDscr.nation	YES
Coverage (Type= Administrative unit)	codeBook.stdyDscr.stdyInfo.sumDscr.geogUnit	
Coverage (Type= Chronological period)	codeBook.stdyDscr.stdyInfo.sumDscr.collDate	
Rights	codeBook.docDscr.citation.prodStmt.copyright or codeBook.stdyDscr.dataAccs.useStmt	
Contact	codeBook.docDscr.citation.distStmt.contact	

Note: DDI is very extensive and some elements of the metadata (e.g., Title) could be defined at “Document Description” level (codebook.docDscr) or at “Study Description” level (codeBook.stdyDscr).

### Technical implementation

#### For Dataverse users

Although Dataverse can collect almost all CG Core elements, special attention should be given to those that depend on a vocabulary or a list. For example, CRPs as a contributor should be a drop down selection to avoid invalid entries. Dataverse users can implement this by customizing their Dataverse installation, however those users using Harvard’s installation at <https://dataverse.harvard.edu> have two options: 1) Institutionalize the usage of vocabularies and lists so the person responsible for uploading the metadata uses the appropriate terms or, 2) move the entire Dataverse from Harvard to a custom installation elsewhere.

#### For users harvesting CG Core e.g., CGIAR System Organization

Dataverse has a robust API thus by implementing the necessary DDI elements it should be straight forward to harvest almost all CG Core for a dataset. The only missing element is “Language” which could be set to “EN” at extraction time.

Because Dataverse elements can be defined at “Document” or “Study” levels software implementing the extraction of the metadata should check if the elements are present in both levels.

### DSPACE repositories

DSPACE repositories like CGSpace use Dublin Core for storing metadata. Since CG Core is based in Dublin Core (with the exception of the element “Contact”) it should be straight forward to extract CG-Core from D-Space repositories, however DSPACE /CGSpace users need to check if the following list of elements are present for each publication:

Element in CG Core	Element in DSpace	Vocabulary / list required
Title	dc.title	
Creator	dc.contributor.author	
Subject	dc.subject and cg.subject.[centre]	YES
Description	dc.description.abstract	
Publisher	dc.publisher	
Contributor	cg.contributor and subelements, dc.description.sponsorship (for funder/sponsor) and cg.identifier.[centre]project	YES
Date (Type=Production)	dc.date.issued	
Date (Type= Availability)	dc.date.available	
Type	dc.type	YES
Format	No mapped to DSpace	YES
Identifier (Type= Identifier)	cg.identifier.url, dc.identifier.uri and cg.identifier.doi	
Identifier (Type= Citation)	dc.identifier.citation	
Source	dc.source	
Language	dc.language.iso	
Relation	dc.relation	
Coverage (Type= Geospatial)	dc.coverage.spatial (Point Encoding Scheme, <a href="http://dublincore.org/documents/dcmi-point/">http://dublincore.org/documents/dcmi-point/</a> )	
Coverage (Type= Region)	cg.coverage.region	YES
Coverage (Type= Country)	cg.coverage.country	YES
Coverage (Type= Administrative unit)	cg.coverage.subregion	
Coverage (Type= Chronological period)	dc.coverage.temporal (Period Encoding Scheme, <a href="http://dublincore.org/documents/dcmi-period/">http://dublincore.org/documents/dcmi-period/</a> )	
Rights	Not in DSpace	
Contact	Not in DSpace	

Note: DSPACE implements different metadata elements at community level; for example CCFAS identify a grant code in “cg.identifier.ccfasproject” while others could implement it in a different element thus the mapping between DSPACE and CG Core could vary.



## Technical implementation

### For DSpace users

DSpace metadata elements can be customized at community level. Special attention should be given to elements that depend on a vocabulary. For example, CRPs as a contributor should be a drop down selection to avoid invalid entries. This can be done when customizing the DSpace schema.

### For users harvesting CG Core e.g., CGIAR System Organization

DSpace has a robust API thus by implementing the necessary elements it should be straight forward to harvest almost all CG Core for a publication with the following exceptions:

Contact could be set to blank at extraction time.

Format could be guessed from the attachment in the Bitstreams.

Rights could be set to a particular CC license at extraction time.

Authors IDs like ORCID are not stored in the DSpace database thus are not included in the API.

Author's affiliation are independent elements in the metadata thus is not possible automatically link an affiliation to an author.

Because DSpace schema elements can vary at community level, software implementing the extraction must be developed to specifically harvest a community.

## CKAN repositories

CKAN does not enforce a particular metadata schema like DDI or Dublin Core. CKAN repository users must write an extension to expand the metadata schema and accommodate for CGCore. The following table shows a possible implementation:

Element in CG Core	Element in DDI	Vocabulary / list required
Title	Implemented in schema as free text	
Creator	Implemented in schema as free text, however it can be re-implemented as tags to accommodate for IDs	
Subject	Implemented as free (not vocabulary) tags. It can be re-implemented as a vocabulary	YES
Description	Implemented as markdown text	
Publisher	Needs implementation by extension in extras	
Contributor	Needs implementation by extension. Lists like CRPs and Centres should be implemented as tag vocabularies while individuals in extras	YES
Date (Type=Production)	Needs implementation by extension in extras	
Date (Type= Availability)	Needs implementation by extension in extras	
Type	Needs implementation by extension as a tag vocabulary	YES
Format	Implemented at resource level. Can be re-implemented at dataset level as a tag vocabulary	YES
Identifier (Type= Identifier)	Needs implementation by extension in extras	
Identifier (Type= Citation)	Needs implementation by extension in extras	
Source	Needs implementation by extension in extras	
Language	Needs implementation by extension in extras or as a tag vocabulary	
Relation	Needs implementation by extension in extras	
Coverage (Type= Geospatial)	Needs implementation by extension in extras	
Coverage (Type= Region)	Needs implementation by extension as vocabulary tags	YES
Coverage (Type= Country)	Needs implementation by extension as vocabulary tags	YES
Coverage (Type= Administrative unit)	Needs implementation by extension in extras	
Coverage (Type= Chronological period)	Needs implementation by extension in extras	
Rights	License is implemented in the schema as a list, however it can be re-implemented in extras	
Contact	Needs implementation by extension in extras	

## Technical implementation

### For CKAN users

CKAN metadata can be easily customized using extensions. Special attention should be given to elements that depend on a vocabulary. For example, CRPs as a contributor should be a drop down selection to avoid invalid entries.

### For users harvesting CG Core e.g., CGIAR System Organization

CKAN has a robust API thus by implementing the necessary elements it should be straight forward to harvest all CG Core for a dataset. However, because the customization can vary from one CKAN user to another, software implementing the extraction must be implemented to specifically harvest a repository.

### Annex I: Example of a Metadata Template

Element	Value	Attributes									
		ID	Type	Affiliation	Vocab	Role	X	Y	Event	Email	
Title	IMPACT Lite - Nyando										
Creator	Silvestri S.			CABI International							
Creator	Quiros C.	0000-0002-9485-9961@ORCID		International Livestock Research Institute							
Creator	Mutie I.			International Livestock Research Institute							
Creator	Ndiwa N.			International Livestock Research Institute							
Creator	N'dungu A.			World Agroforestry Centre							
Creator	Rufino M.			Lancaster University							
Creator	Herrero M.	0000-0002-7741-5090@ORCID		Commonwealth Scientific and Industrial Research Organisation							
Creator	Kiplimo J.			International Livestock Research Institute							
Subject	climate change				GACS						
Subject	farming systems				GACS						
Subject	food security				GACS						
Description	The Integrated Modelling Platform for Mixed Animal Crop systems (IMPACT) was developed to encourage data sharing by using standard protocols, and allowing tools to be linked to facilitate evaluations of various farming systems. There was however a need to further improve the tool, to make it easier and more effective to use, as it took considerable time to complete an interview. With this in mind, CCAFS ( <a href="http://ccaafs.cgiar.org/">http://ccaafs.cgiar.org/</a> ) commissioned the International Livestock Research Institute (ILRI) the task to redesign IMPACT into a lighter tool for household characterization (IMPACT Lite). IMPACT Lite helps capture the diversity of farming activities and characterize the main agricultural production systems. It is really useful to anyone with the ambition to better understand farmers' production systems and their dynamics.										
Publisher	International Livestock Research Institute										
Contributor	CGIAR Research Program on Climate Change,		CRP								



### 30. CG Core Basic for Researchers (Excel)

- *Data Management*
  - *Archiving*
  - *Maintenance*
- *Communications*
  - *Data dissemination/publicity*
- *Using the Data*
  - *User manuals/documentation*
- *Main*
- *When*
  - *Management of research processes*
  - *Delivery of research products*
- *PI, Researcher*

Please find below the document

## CG Core Basic for Researchers

DC Element	Qualifier	Required?	Definition
Title	<b>Title of resource</b>	Required	Official or unofficial title of the document, data set, image, etc.
Creator	<b>Name of resource creator</b>	Required	Creators of the item—typically a person. Could be an organization in case of corporate authors (e.g. Center reports)
Creator	<b>ID of resource creator - if any</b>	Required when applicable	ID of creator; use if ORCID, SCOPUS, or other type of creator ID scheme is in use
Creator	<b>ID type of resource creator - if any</b>	Required when applicable	Used to indicate the type of Creator ID – ex: SCOPUS, ORCID, etc.
Subject	<b>General subject matter</b>	Required	Subject matter of the research, technologies tested, etc.
Subject	<b>AGROVOC subject term</b>	Optional	AGROVOC subject matter or research area
Subject	<b>Subject - other vocabularies (e.g. MeSH)</b>	Required if applicable	Subject matter or research area from domain-specific vocabularies, if missing from AGROVOC
Description	<b>Abstract of work</b>	Required	Abstract or other description of the item
Publisher	<b>Publisher of journal</b>	Required when applicable	Entity responsible for publication, distribution, or imprint
Contributor	<b>CGIAR Center name</b>	Required	Research Centers with which creator(s) are affiliated
Contributor	<b>non-CGIAR entity name</b>	Required when applicable	Non-CGIAR partner entity with which creator/s are affiliated
Contributor	<b>CRP</b>	Required when applicable	CGIAR Research Program with which the research is affiliated
Contributor	<b>Funding agency</b>	Required	Funder, funding agency or sponsor
Contributor	<b>Project</b>	Required	Name of project with which the research is affiliated
Date	<b>Publication or creation date</b>	Required	Publication, creation, end of trial, or issue date
Date	<b>Embargo date for publication</b>	Required when applicable	Used when an item has an embargo by publisher (ex: 6 or 12-month embargo)

DC Element	Qualifier	Required?	Definition
Type	<b>Type of resource</b>	Required	Nature or genre of item/content; e.g., article, book chpt, poster, data set, audio etc
Format	<b>File format</b>	Required	File format of item e.g.: PDF; jpg etc.
Identifier	<b>Unambiguous identifier of resource</b>	Required when applicable	Unambiguous reference to resource such as doi, uri
Identifier	<b>Citation</b>	Required when applicable	Human-readable, standard bibliographic citation for the item
Source	<b>Journal / Proceedings title</b>	Required when applicable	Journal/conference title; vol., no. (year)
Language	<b>Language</b>	Required	Language of the item; use ISO 639-1 (alpha-2) or ISO 639-2 (alpha-3).
Relation	<b>Files related to resource</b>	Optional	Supplemental files, e.g. data sets related to publications or larger “whole” (book chapters etc)
Coverage	<b>Region</b>	Required	Supra-national areas (above country level) related to the item being described
Coverage	<b>Country</b>	Required	Country/countries related to the data which was collected in resource
Coverage	<b>Admin unit - level 1</b>	Required when applicable	Sub-national administrative areas such as provinces, states, or districts
Coverage	<b>Geospatial coordinates</b>	Required	Coordinates or polygon points for boundaries of area where research was conducted (in decimal degrees)
Coverage	<b>Start date of activity</b>	Required	Chronological period: start date of activity described in resource
Coverage	<b>End date of activity</b>	Required	Chronological period: end date of activity described in resource
Rights	<b>Terms of use</b>	Required	Rights, licensing, IPR, or permission statement
Contact	<b>Point of contact</b>	Optional	For data: email address for group or department to contact in case of questions

*Please see accompanying document before using this template*

### 31. CG Core Basic for Researchers

- *Data Management*
  - *Archiving*
  - *Maintenance*
- *Communications*
  - *Data dissemination/publicity*
- *Using the Data*
  - *User manuals/documentation*
- *Main*
- *When*
  - *Management of research processes*
  - *Delivery of research products*
- *PI, Researcher*

Please find below the document

# CG Core Metadata for Researchers

## Introduction

This document accompanies the Excel template of the same name. The template contains elements based on the Dublin Core Metadata Schema. We also refer you to the CG Core Metadata Schema and Application Profile which provides a more in-depth description of the elements of the schema.

Note that this document and the accompanying template, describe what we have referred to in the Introduction to Metadata document, as the Study Catalogue. In addition to the Study Catalogue you would also need to supply a data dictionary for each dataset you produce.

## Title

This is a required element and is typically the name by which the resource is formally known. For example, "Managing for timber and biodiversity in the Congo basin".

## Creator

### Name of resource creator

This is another required element and is the name of the person, organisation or service that created the resource. This is a multiple element as all persons and/or organisations responsible for the resource should be included. Examples: "Garlick, CA", "International Centre for Tropical Agriculture (CIAT)".

### ID of resource creator

This element is only required if it is available. This might be the ORCID (Open Researcher and Contributor ID) of the creator, SCOPUS Author Identifier, or other type of creator ID scheme if in use. For information about ORCID see the website at <https://orcid.org> ; for information about SCOPUS go to <https://www.elsevier.com/solutions/scopus>

### ID type of resource creator

If an ID of the resource creator has been used above; e.g. the creator's ORCID has been used, then for this element you should specify the type of resource; e.g. ORCID, SCOPUS, etc.

## Subject

### General Subject matter

This is a required element and is the subject matter of the research, technologies tested, crops involved in the research, methodologies, etc. The format should be single words or short phrases and where possible controlled vocabularies should be used. Examples: "Cattle", "Dairy", "Maize".

### AGROVOC subject term

AGROVOC subject matter or research area. AGROVOC is a multilingual controlled vocabulary covering all areas of interest of the Food and Agriculture Organisation of the United Nations (FAO). For details



see the website at <http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus>

## Subject – other vocabularies

This element is the subject matter or research area from domain-specific vocabularies if it is missing from AGROVOC. An example of another vocabulary is MeSH (Medical Subject Headings) – see the website at <https://www.ncbi.nlm.nih.gov/mesh> for further details.

## Description

The description is required and is generally the abstract or other description of the item. This is especially important for datasets, software, journal articles, working papers, reports and other types of written materials. Descriptions can be provided in multiple languages if appropriate and available.

## Publisher

This is the entity responsible for making the resource available. This could be a person, organisation or a service. Note this is not the journal title, but the publisher of the journal. This element is required when applicable.

## Contributor

### CGIAR Centre name

This is the Research Centre(s) with which the creator(s) are affiliated. There may be multiple contributor elements. For example: “International Centre for Tropical Agriculture (CIAT)”, “World Agroforestry Centre (ICRAF)”.

### Non-CGIAR entity name

If the creators are affiliated to non-CGIAR partners, then these should be mentioned here. For example: “Statistics for Sustainable Development (Stats4SD)”.

### CRP

If the research is linked to one or more CGIAR Research Programmes, then these should be mentioned here. For example: “CGIAR Research Programme on Climate Change, Agriculture and Food Security (CAAFS)”.

### Funding Agency

This is a required element and you should give the name of the funder, funding agency or sponsor.

### Project

This is also a required element and should be the name of the project with which the research is affiliated.

## Date

### Publication or Creation Date

This should be the date when the resource was created in its final form ready for publication. The date should be in one of the following formats: YYYY-MM-DD, YYYY, YYYY-MM.

## Embargo Date for Publication

This is required if there is an embargo on publication for some reason. This would be the date that the resource will become available.

## Type

This is a required element and is the type of the resource. Use singular words or phrases – e.g. “Image”. Terms should be taken from the following list which can be expanded to include other types of research outputs and information products, based on the contents in a particular repository:

- Audio
- Book
- Book Chapter
- Dataset
- Extension Material
- Image
- Map
- Model
- Peer-reviewed journal article
- Policy Brief
- Report
- Software
- Source Code
- Thesis
- Training Material
- Video

## Format

A required element identifying the format of the resource. A list is available in Wikipedia at [https://en.wikipedia.org/wiki/Media\\_type](https://en.wikipedia.org/wiki/Media_type). Some repositories will pull this information from the object automatically while others require manual input for this element. For example: “application/pdf”, “image/jpeg”.

## Identifier

### Unambiguous identifier of resource

This is the reference to the resource which might be the DOI (Digital Object Identifier) or URI (Uniform Resource Identifier). When datasets are created within Dataverse for example, a unique identifier is automatically generated for the resource.

## Citation

This is the standard, human-readable bibliographic citation for the resource.

## Source

This is the original journal article or other type of material where an item was originally published. This is used for journal articles, data articles, conference proceedings, etc. Examples: “Journal of Development Economics”, “World Development”.

## Language

This is a required element and is the language of the item. This is for human languages only, not computer/software programming languages. Use ISO 639-1 (alpha-2) or ISO 639-2 (alpha-3). For example: “EN” (English), “ES” (Spanish), “FR” (French). See Wikipedia [https://en.wikipedia.org/wiki/List\\_of\\_ISO\\_639-1\\_codes](https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes) for a list of language codes.

## Relation

These are supplemental files, e.g. data sets related to publications; related to the resource. For example: “<https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/CCAFS-OBS-2012>”

## Coverage

### Region

This is the supra-national area (i.e. above country level) related to the item being described. Example: “Eastern Africa”.

### Country

Country or countries related to the data which was collected in the resource. Use the country names from ISO 3166 (<https://www.iso.org/obp/ui/>) Example: “Kenya”. Note: If there is a lack of clarity regarding countries, best practice is not to include a country. Likewise, in cases where research has taken place in politically sensitive areas and where including country-level information could be problematic, it is best not to include such details in the record.

### Admin unit – level 1

Sub-national administrative areas such as provinces, states or districts. Example: “Makueni”. Note: in order to facilitate discovery, all records (where applicable) should include tagging of information at sub-national level (ideally district level).

### Geospatial coordinates

This should be the coordinates or polygon points for boundaries of the area where the research was conducted. These should be recorded in decimal degrees.

### Start date of activity

Start date of the activity described in the resource. For example, when running a survey this would be the start date of the fieldwork.

### End date of activity

The end date of the activity described in the resource. These two elements give the timing and length of the fieldwork related to the resource.

## **Rights**

These are the terms of use for the resource and should include any licensing restrictions and IPR. For example: "Access (unrestricted); Re-use (non-commercial only); Permissions (Centre e-mail)"

## **Contact**

This is the point of contact for information about the resource. This should be the email address for the group or department to contact in case of questions.

# Archiving and Sharing

## Main

### 32. Principles for Archiving and Sharing

- *Data Management*
  - *Archiving*
- *Strategic Planning*
  - *Data Ownership*
- *Using the Data*
  - *Data Protection, privacy, anonymisation*
  - *Sharing and Access to data*
- *Legal*
  - *Open Access/restrictions*
  - *Ownership of data*
- *Main*
- *When*
  - *Decisions while designing*
  - *Management of research processes*
  - *Delivery of research products*
- *PI, Researcher*

Please find below the document

# Principles for Archiving and Sharing

## Introduction

This guide explains some of the benefits of archiving research data. We acknowledge intellectual property and highlight ethical aspects such as anonymity and confidentiality of information providers.

## Why should I archive my Data?

In the past, and still today in some cases, researchers were reluctant to share their data. This might be because they fear someone else might get the credit for the work they have done, or perhaps they want more time to carry out their own analyses.

However, most funding bodies are now viewing research data as a public resource and, as a condition of funding, are insisting that data are put into the public domain in a timely fashion. Issues of intellectual property, confidentiality, etc., have been, and are continuing to be addressed.

CGIAR has adopted an Open Access policy with respect to Research Data. This means that all research data, generated as a result of research funded by CGIAR programmes, must, subject to confidentiality of respondents, be deposited in a suitable repository and made publicly available as soon as possible. Details of the CGIAR Open Access policy are available [here](#). A copy of the policy is also available as part of this pack.

## Advantages of Sharing Data

There are many advantages to sharing data. These include:

- Encouraging scientific enquiry and debate;
- Promoting innovation and potential new uses of the data;
- Developing new collaborations between researchers;
- Maximising transparency and accountability;
- Encouraging the improvement and validation of research methods;
- Eliminating the need to collect the same data again, thus reducing time and costs;
- Increasing the impact and visibility of the research.

Remember that sharing data is a two-way process; while other researchers are able to make use of your research data, you are also able to make use of data shared by others.

## Principles

General principles for archiving data might include:

- Publicly funded research data are a public good which should be made openly available in a timely and responsible manner;
- Sufficient metadata should be recorded and made available to enable others to understand the research and the potential re-use of the data;
- Researchers should have a limited period of privileged use of the data collected to enable them to publish the results of their research – the time might vary but is expected to be no

more than 12 months after data collection or within 6 months of publication, whichever is the sooner;

- All users of the data should acknowledge the sources of their data and abide by the terms and conditions under which they are accessed.

## Intellectual Property

When data are archived, the Intellectual Property or Copyright remains with the researcher(s); it does not transfer to the hosting organisation. For example, the CCAFS Baseline Surveys are archived in Dataverse and hosted at Harvard. The Dataverse Project acts as a publisher in this instance but does not have any rights over the data collections it houses. Users of research data should acknowledge the source of their data, and it is therefore useful for data creators to specify in the archive how they would like to be acknowledged. Dataverse, for example, creates a unique citation for each dataset, and part of the terms and conditions of use is that this citation is used in scholarly references.

The guide on Data Ownership and Authorship includes a short section on Rights and Responsibilities which includes Intellectual Property.

## Anonymity and Confidentiality

Before archiving data, you should ensure that the dataset is anonymised – i.e. an individual cannot be identified from their data. Obviously, this would include removing names and addresses of individuals, but there are other things to consider. Anonymising data can be time-consuming, so ensure it is adequately planned for.

### Quantitative Data

Techniques for anonymising quantitative data may involve removing or aggregating variables or reducing the precision of a variable.

#### *Remove Direct Identifiers*

Direct identifiers include names, addresses, telephone numbers, etc. These are generally not needed for secondary research but are collected for checking purposes or to enable follow-up. These variables can easily be replaced by a code in the data.

#### *Aggregate or reduce the precision of a variable*

Examples here might include recording just the year or the year and month of birth rather than the full birth date. Detailed geo-references could be problematic as they could identify individuals. They could be replaced by alternative variables that typify the geographical position; e.g. poverty index, population density, altitude, vegetation type, etc. This would maintain the value of the data without disclosing the exact locations.

#### *Restrict the upper or lower ranges of a continuous variable*

If the values for an individual are unusual within the wider group researched, you could collapse unusually large or small values into a single code. For example, a top code of “50 hectares or more” could be applied to land ownership.

### Qualitative Data

When anonymising qualitative material such as transcribed interviews, identifiers should not be crudely removed as this can distort the data. Instead we suggest using pseudonyms, replacement

terms, or vaguer descriptions. The aim is to achieve a reasonable level of anonymisation whilst maintaining maximum content.

Suggestions for the anonymisation of text include:

- 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 – e.g. by 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 anonymised data file. The following table gives an example of such a log file:

INTERVIEW NUMBER	ORIGINAL VALUE	CHANGED TO
1	Age 54	Age range 50-55
1	20 <sup>th</sup> June	June
1	Cathy (real name)	Jane (pseudonym)
2	Station Hill Primary School	A primary school
2	Rachel	My friend

Anonymising 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.

## Consent

Informed consent is an ethical requirement for most research and must be considered and implemented throughout the research lifecycle from planning through to publication and archiving. Gaining consent must make provision for sharing data.

Researchers should inform participants about how the data they are collecting 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 should 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, and contact details for researchers.

You need to consider the type of data that you will be collecting and whether you intend to follow-up the individual at a later date. Bear in mind that a respondent may be happy to participate initially but may not want to be involved in any follow-ups.



You will need a consent form, and this should allow the participant to clearly respond to each of the following points:

- They have read and understood information about the project;
- They have been given the chance to ask questions;
- They voluntarily agree to participate in the study;
- They understand that they can withdraw at any time without giving a reason;
- They understand that they can refuse to answer one or more of the questions;
- They understand how the data are to be used and archived.

There should also be separate consent sought for use of any audio/visual data such as recordings, videos or photos – some might agree to complete a questionnaire but might not want photos of themselves to be made public.

If your research involves working with children, then consent must be sought from the parent/guardian as well as from the child.

## What should I archive?

Of course, when we archive data it is not just the data file itself that we archive. Many data files are of limited use without the accompanying documentation. At the minimum your archive should include:

- The Activity Protocol so others can clearly see the focus of your research;
- The Data Management Plan to show the steps you intended to follow to ensure high quality data;
- The Data Entry system if one has been used;
- The Fieldworker Manual which will detail the procedures used to collect the data;
- A blank copy of the Questionnaire – adding variable names to the questionnaire would be useful for interpreting the data;
- The Data Quality Report which would highlight any problem areas in the data and give suggestions for their use;
- The Metadata Document used to describe the data;
- Etc.

This pack includes a separate checklist of data and documents to submit for archiving.

## Preparing for Archiving – Using a DDS

We strongly recommend starting to prepare your archive early in the project lifecycle; if all the preparation is left to the end (as is often the case) then you will find yourself struggling to pull together all the documents and information you need as many of those with the required information will have moved on to other projects. Consider using a Data and Document Storage facility (DDS) – see the separate document in this pack on Data and Document Storage for further information. By using a DDS, you can start to build your archive from the start of the project. When you are ready to archive, it will be much quicker and easier to transfer the files across to the repository.

## Summary

There are clear benefits to archiving and sharing research data but there are also responsibilities. You must ensure the confidentiality of your respondents, ensure you have informed consent from respondents and ensure your data have been anonymised.

Think of data sharing as a two-way process – if you are not willing to share your data with others, you cannot expect others to share their data with you.

## References

- CGIAR Open Access and Data Management Policy -  
<https://cgspace.cgiar.org/bitstream/handle/10947/2875/CGIAR%20OA%20Policy%20-%20October%20202013%20-%20Approved%20by%20Consortium%20Board.pdf?sequence=4>
- UK Data Archive -  
<https://www.ukdataservice.ac.uk/manage-data/legal-ethical/consent-data-sharing/consent-forms>
- The Dataverse Project -  
<https://dataverse.org/>

## Associate Videos

Videos accompanying the original release of the CCAFS Data Management Support Pack in 2013 are available as a playlist on the Statistical Services Centre YouTube Channel at <https://www.youtube.com/channel/UCs7EU95YMjvNozJKCD92xQ/playlists>. These videos have not been updated since the original release but are mostly still relevant.

In particular the playlist includes a video on “Archiving and Sharing” which is available at: <https://www.youtube.com/watch?v=H8sO21P5RBc&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpj&index=10>

### 33. Data and Documents to Submit for Archiving – a checklist

- *Data Management*
  - *Archiving*
- *Main*
- *When*
  - *Management of research processes*
  - *Delivery of research products*
- *Researcher, Technician*

Please find below the document

## Data and Documents to Submit for Archiving – a Checklist

### Introduction

The following is provided as an example of what files should be submitted to the data archive. We suggest you adjust this list to suit your own research project.

Document	Archived
Copy of the Data Ownership Agreement	<input type="checkbox"/>
Activity Protocol	<input type="checkbox"/>
Data Management Plan	<input type="checkbox"/>
Data Entry System (if used)	<input type="checkbox"/>
Fieldwork Manual	<input type="checkbox"/>
Questionnaires (if used)	<input type="checkbox"/>
Analysis Plan	<input type="checkbox"/>
Data Quality Report	<input type="checkbox"/>
Raw Data (anonymised version)	<input type="checkbox"/>
Primary Data (anonymised)	<input type="checkbox"/>
Metadata document	<input type="checkbox"/>
Analysis Program (R code, SPSS syntax, etc.)	<input type="checkbox"/>
Analysis Output	<input type="checkbox"/>
Interim Reports	<input type="checkbox"/>
Final Report	<input type="checkbox"/>

#### 34. Data and Documents to Submit for Archiving – a checklist

- *Data Management*
  - *Archiving*
- *Main*
- *When*
  - *Management of research processes*
  - *Delivery of research products*
- *Researcher, Technician*

Please find below the document

# Portals for Archiving and Sharing

## Introduction

This guide distinguishes between repositories for data and those for publications. Although there are numerous repositories available, the ones used most often by CIAT and CCAFS are Dataverse for archiving data and related documents, and DSpace for storing publications. The guide notes the difference between “reports” which should be included with the data archive, and peer-reviewed publications.

## Data Repositories

We start by looking at the main data repositories for CIAT and CCAFS data and related documentation. Note that data reports, fieldwork reports, interim and final analysis reports based on the analysis plan, should be archived along with the data to a data repository.

## Dataverse

Dataverse is an open source web application for sharing, preserving, exploring and analysing research data. It was created by the [Institute for Quantitative Social Science](https://dataverse.org/) at Harvard University. A separate document in this pack describes Dataverse in more detail. The website for the Dataverse Project is <https://dataverse.org/>

### CIAT Dataverse

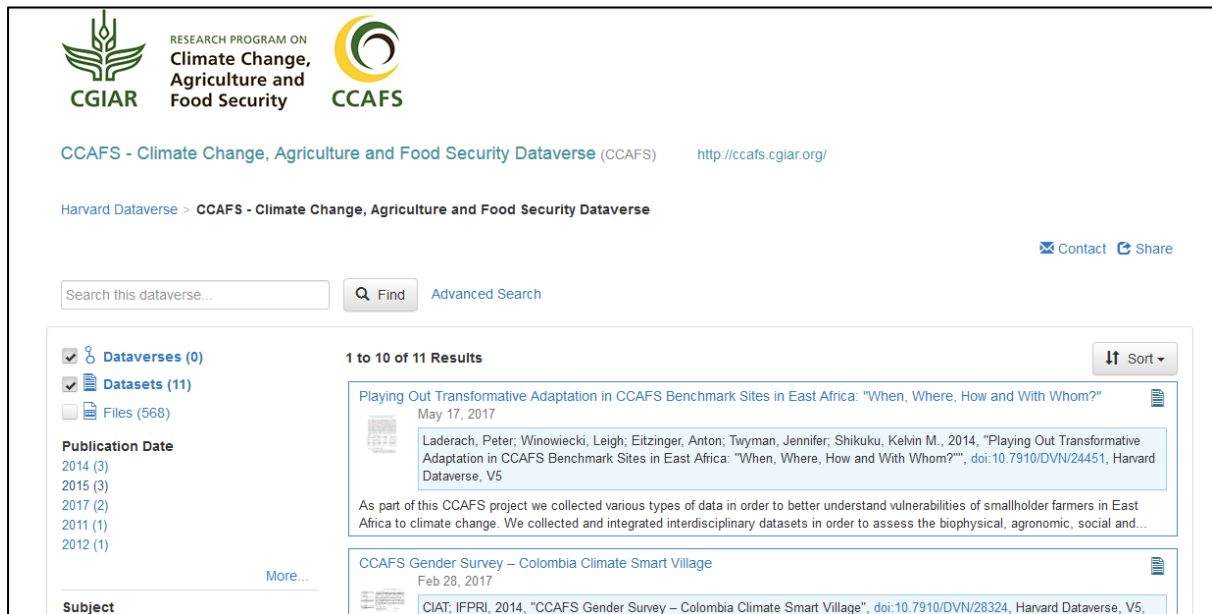
The CIAT Dataverse can be found at the following link: <https://dataverse.harvard.edu/dataverse/CIAT> and the image below shows the front page of this Dataverse.

The screenshot displays the CIAT Dataverse website. At the top, the CIAT logo is on the left, and the text 'International Center for Tropical Agriculture Since 1967 Science to cultivate change' is in the center. To the right, a world map is partially visible with the text 'CIAT Research Online' in red. Below the header, there are two lines of text: 'CIAT - International Center for Tropical Agriculture Dataverse (CGIAR)' and 'CIAT - Eco-efficient agriculture for the poor'. A breadcrumb trail shows 'Harvard Dataverse > CIAT - International Center for Tropical Agriculture Dataverse'. On the right, there are links for 'Contact' and 'Share'. A paragraph of text describes CIAT's mission and the availability of research data. Below this, three boxes represent different dataverses: 'Agrobiodiversity Dataverse', 'Decision and Policy Analysis Dataverse', and 'Soils Dataverse'. A search bar is present with a 'Find' button and a link to 'Advanced Search'. The search results section shows '1 to 10 of 118 Results' and a 'Sort' dropdown. The first result is 'Climate Risk Vulnerability Assessment to inform sub-national decision making in Vietnam' by Parker, Louis; Bourgoin, Clement; Martinez-Valle, Armando; Läderach, Peter, dated Feb 23, 2018. The second result is 'CLEANED X - Version 1.0.1' by Notenbaert, An; Birthe, Paul; Mukiri, Jessica; Birnholz, Celine; Koge, Jessica, dated Jan 31, 2018.

At the time of writing, the CIAT Dataverse houses 118 datasets which are divided among 3 sub-Dataverses: Agrobiodiversity, Decision and Policy Analysis, and Soils. The most recent dataset was published on the Dataverse on 23<sup>rd</sup> February 2018.

### CCAFS Dataverse

The CCAFS Dataverse can be found at the following link: <https://dataverse.harvard.edu/dataverse/CCAFSbaseline> and the image below shows the front page of this Dataverse.



There are currently 11 datasets housed on the CCAFS Dataverse which includes the Households Baseline, Village Baseline and Organisational Baseline Studies.

## AgTrials

AgTrials is the Global Agricultural Trial Repository and Database. This is an information portal developed by CCAFS which provides access to a database on the performance of agricultural technologies at sites across the developing world. It builds on decades of evaluation trials, mostly of varieties, but includes any agricultural technology for developing world farmers.

AgTrials is available at the following link: <http://agtrials.org/> and the front page of the site is shown below. With the interface you can:

- Share data and information on evaluations of agricultural technology;
- Acquire agricultural evaluation datasets for your own research;
- Explore the geographic dimensions of agricultural evaluation.

The screenshot shows the AgTrials website homepage. At the top, the logo 'AgTrials' is displayed next to the text 'The Global Agricultural Trial Repository and Database'. To the right are logos for CGIAR and CCAFS, along with the text 'RESEARCH PROGRAM ON Climate Change, Agriculture and Food Security'. A green navigation bar contains links for Home, About Us, Trial, Statistics, and Contact Us, along with Sign In and Sign Up buttons. Below the navigation bar are three main action buttons: 'Search Trials' (with a magnifying glass icon), 'Add New Trial' (with a plus icon), and 'Upload Batch of Trials' (with an upload icon). Each button has a descriptive subtitle below it: 'Simple and Advanced Searches', 'Upload your data to AgTrials', and 'Upload your data in batch mode' respectively. The main content area features a section titled 'What is AgTrials?' with a paragraph explaining the portal's purpose. To the right of this text is a world map showing the number of trials in various regions, with a legend indicating trial counts from 1 to 10,001+. The map data is as follows:

Region	Number of trials
North America	4,734
South America	1,180
Europe	1,752
Asia	6,325
Africa	3,063
Oceania	2,521
Indian Ocean	3,022
South America	13,623

Below the map is a section titled 'What you can do' with three dark blue boxes containing the following text:

- Share data and information on evaluations of agricultural technology.
- Acquire agricultural evaluation data sets for your own research.
- Explore the geographic dimensions of agricultural evaluation.



## CCAFS-Climate

The CCAFS-Climate data portal provides global and regional future high-resolution climate datasets that serve as a basis for assessing the climate change impacts and adaptation in a variety of fields including biodiversity, agricultural and livestock production, and ecosystem services and hydrology.

CCAFS-Climate can be found at: <http://www.ccafs-climate.org/> and the image below shows the front page of the site. The site includes a video tutorial showing how to download data from the site.



The screenshot shows the homepage of the GCM Downscaled Data Portal. The header features the title 'GCM DOWNSCALED DATA PORTAL' on the left, and navigation links for 'Contact' and 'About Us' on the right. Below the title are logos for CGIAR, the Research Program on Climate Change, Agriculture and Food Security, and CCAFS. A main navigation menu includes 'Home', 'Data', 'Methods', 'Documentation', 'Links', and 'Citations'. Below this is a row of six vertical panels: 'Data' (with a globe), 'Methods' (with a map), 'Useful Documents' (with a document icon), 'Links' (with a globe), 'Citations' (with a data table), and 'Contact' (with a mailbox). Below the panels, there is a section titled 'Data Provided by the CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS)' followed by a disclaimer and a Creative Commons Attribution-NonCommercial 4.0 International License logo.

## Publication Repositories

By “publications” in this context we are referring to peer-reviewed publications and articles such as those that might be submitted to journals. These publications should, wherever possible, include a reference to the location of the data used for any analysis within the publication. This adds credence to the results as it is possible for other researchers to download the relevant data and confirm the results.

## DSpace

The main publication repository used by CIAT and CCAFS is DSpace - <http://www.dspace.org/>. DSpace open source software is a repository application used by more than 1000+ organisations and institutions worldwide; it provides durable access to digital resources. DSpace preserves and enables easy and open access to all types of digital content including text, images, moving images and datasets.

## 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. It is a collaboration of several centres, the CGIAR system management office and research programmes. It is hosted by the International Livestock Research Institute (ILRI).

The screenshot shows the CGSpace website interface. At the top, there is a green header with the CGIAR logo and the text 'CGSpace A Repository of Agricultural Research Outputs'. A search bar is located in the top right corner. Below the header, the main content area is divided into several sections. On the left, there is a 'Welcome' section with a brief description of the repository. In the center, the 'Communities in CGSpace' section is displayed, listing various research programs and their collection counts. Two items are highlighted with a yellow background: 'CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS) [3450]' and 'International Center for Tropical Agriculture (CIAT) [12190]'. On the right side, there is a sidebar with 'My Account' (Login, Register), 'Discover' (Authors), and a list of contributing organizations with their respective collection counts.

**Communities in CGSpace**

Select a community to browse its collections.

- Africa RISING [964]
- AgriFood Chain Toolkit [103]
- Animal Genetic Resources Virtual Library [1055]
- CGIAR Collective Action in Eastern and Southern Africa [47]
- CGIAR Global Mountain Program [8]
- CGIAR Platform for Big Data in Agriculture [1]
- CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS) [3450]**
- CGIAR Research Program on Dryland Systems [122]
- CGIAR Research Program on Integrated Systems for the Humid Tropics [115]
- CGIAR Research Program on Livestock [220]
- Feed the Future Sustainable Intensification Innovation Lab [19]
- IGAD Livestock Policy Initiative [42]
- International Center for Agricultural Research in the Dry Areas (ICARDA) [164]
- International Center for Tropical Agriculture (CIAT) [12190]**
- International Institute of Tropical Agriculture (IITA) [2336]
- International Livestock Research Institute (ILRI) [15857]

**My Account**

- Login
- Register

**Discover**

**Authors**

- Technical Centre for Agricultural and Rural Cooperation (11981)
- International Center for Tropical Agriculture (1884)
- International Livestock Research Institute (1739)
- CGIAR Research Program on Climate Change, Agriculture and Food Security (977)
- CGIAR Secretariat (656)
- CGIAR Consortium Office (624)
- Grace, Delia (604)
- International Water Management Institute (391)
- CGIAR Technical Advisory Committee (318)

**Output types**

- Journal Article (15187)
- News Item (11243)
- Report (6805)
- Conference Paper (5791)
- Book Chapter (5279)
- Internal Document (3872)
- Book (3118)

The image above shows the front page of CGSpace - <https://cgspace.cgiar.org/> and we have highlighted both CCAFS and CIAT in the list of Communities in CGSpace.

The image below shows the CCAFS site within CGSpace which currently houses more than 3,400 items. On the site you can search the collection by author, by date, by output type, by region, etc. This makes the site incredibly flexible.

CGSpace  
A Repository of Agricultural Research Outputs

CGSpace Home / CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS)

## CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS)

Search within this community and its collections:

The CCAFS DSpace is a repository of research publications and products from the CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS).  
Visit the CCAFS publications library on the CCAFS website for even more.

**View outputs by CCAFS research flagship:**

- Priorities and Policies for CSA
- Climate Services and Safety Nets
- Low Emissions Development
- Climate-Smart Technologies and Practices
- Gender and Social Inclusion
- Regional Socio-Economic Scenarios
- Data and Tools for Analysis and Planning
- Climate Change and Social Learning

**View outputs from CCAFS Regions:**

- East Africa
- West Africa
- South Asia
- Latin America
- Southeast Asia

Subscribe to RSS Feed  
Subscribe to email alerts for new publications

**Sub-communities within this community**

CCAFS Research Themes and Topics [22]

**Collections in this community**

CCAFS AgClim Letters [39]  
CCAFS Baseline Surveys [82]

**My Account**  
Login  
Register

**Discover**

**Authors**

CGIAR Research Program on Climate Change, Agriculture and Food Security (975)  
Gonsalvez, Julian (125)  
Thornton, Philip (121)  
Kristjansson, Patti (86)  
Vermeulen, Sonja (71)  
Janis, Andy (68)  
Zougmore, Robert B. (65)  
Forch, Wiebke (60)  
Aggarwal, Pramod K. (53)  
Herrero, Mario (50)  
... View More

**Author affiliations**

CGIAR Research Program on Climate Change, Agriculture and Food Security (1202)  
International Center for Tropical Agriculture (171)  
International Livestock Research Institute (129)  
World Agroforestry Centre (65)  
Bioversity International (44)  
Food and Agriculture Organization of the United Nations (33)  
International Maize and Wheat Improvement Center (32)  
International Potato Center (32)  
Wageningen University and Research Centre (30)  
International Institute of Tropical Agriculture (27)  
... View More

**Date Issued**

2010 - 2018 (3433)  
2004 - 2009 (17)  
Output types

The CIAT site within CGSpace has a similar structure as can be seen from the image below:

CGSpace  
A Repository of Agricultural Research Outputs

CGSpace Home / International Center for Tropical Agriculture (CIAT)

## International Center for Tropical Agriculture (CIAT)

Search within this community and its collections:

CIAT Research Online is CIAT's official Open Access repository of products and publications. In this collection, you will find peer-reviewed journal articles, policy briefs, corporate publications, books, book chapters, manuals, working papers, posters, infographics, videos, web tools, presentations and more. This collection is also available through CIAT's website.

**View outputs by CIAT Subject.**

**View outputs from CIAT Regions:**

- Africa
- Asia
- Latin America and the Caribbean

**View outputs by CIAT Research Area:**

- CIAT Agrobiodiversity
- CIAT Decision and Policy Analysis - DAPA
- CIAT Soils

Subscribe to RSS Feed  
Subscribe to email alerts

Enter your email address:  
Subscribe

**Sub-communities within this community**

CIAT Genetic Resources [577]  
CIAT Research Areas [1113]

**My Account**  
Login  
Register

**Discover**

**Authors**

International Center for Tropical Agriculture (1823)  
Rao, Idupulapati M. (229)  
Tohme, J.M. (216)  
Debouck, Daniel G. (200)  
Ceballos, Hernán (155)  
Blair, Matthew W. (145)  
Bellotti, Anthony C. (141)  
Janis, Andy (140)  
Beebe, Stephen E. (133)  
Peters, M. (126)  
... View More

**Author affiliations**

International Center for Tropical Agriculture (1906)  
Pan-Africa Bean Research Alliance (82)  
International Livestock Research Institute (79)  
CGIAR Research Program on Climate Change, Agriculture and Food Security (63)  
World Agroforestry Centre (49)  
International Institute of Tropical Agriculture (41)  
Wageningen University and Research Centre (35)  
International Potato Center (28)  
Universidad Nacional de Colombia (21)  
Bioversity International (19)  
... View More

**Date Issued**

2010 - 2018 (3346)  
2000 - 2009 (3169)  
1990 - 1999 (2648)  
1980 - 1989 (2158)  
1970 - 1979 (845)  
1965 - 1969 (23)

**Output types**

Journal Article (2094)  
Book Chapter (1999)





## CIAT Library Resources

In addition to the CIAT site within CGSpace, CIAT also has a set of library resources available at the following link: <http://ciat.cgiar.org/publications/ciat-library-resources/> This site provides an overview of CIAT Library resources and provides links to Open Access Research Repositories, to CIAT Repositories, and useful guides and newsletters.

# CIAT Library Resources

CIAT Library supports CIAT scientists and staff through the entire research cycle – from finding articles, to publishing, to archiving work, to tracking bibliometric impact.

From this site, we are pleased to provide you with an overview of CIAT Library resources. For more information, please contact Elizabeth Campillo at [CIAT-library@cgiar.org](mailto:library@cgiar.org).

-  Open Access at CIAT
-  CIAT Library resources
-  Data, info and KM blog
-  Subscribe to our newsletter

[Contact us](#)

### Research Databases and E-Journal Subscriptions

Paid research database subscriptions and e-journals can be accessed via our intranet, **CIATNet**. These resources include:

- Web of Science (ISI Thompson Reuters) – indexing and abstracting
- Springer Link – 1500 full-text journals
- ScienceDirect (Elsevier) – full-text access via the Library
- Various e-journal subscriptions – full-text access to 233 scientific journals

Visit these resources on CIATNet [here](#).

### Open Access Research Repositories

### 35. Introduction to Dataverse

- *Data Management*
  - *Archiving*
  - *Data Storage*
- *Technical*
  - *Choice of in-house vs external services/development*
  - *IT Systems (hardware, software, services)*
- *Using the Data*
  - *Dashboards/data displays*
  - *Sharing and Access to data*
  - *User manuals/documentation*
- *Communications*
  - *Data dissemination/publicity*
- *Legal*
  - *Licensing of data and research*
  - *Open access/restrictions*
- *Reference*
- *When*
  - *Management of research processes*
  - *Delivery of research products*
- *PI, Researcher, Technician*

Please find below the document

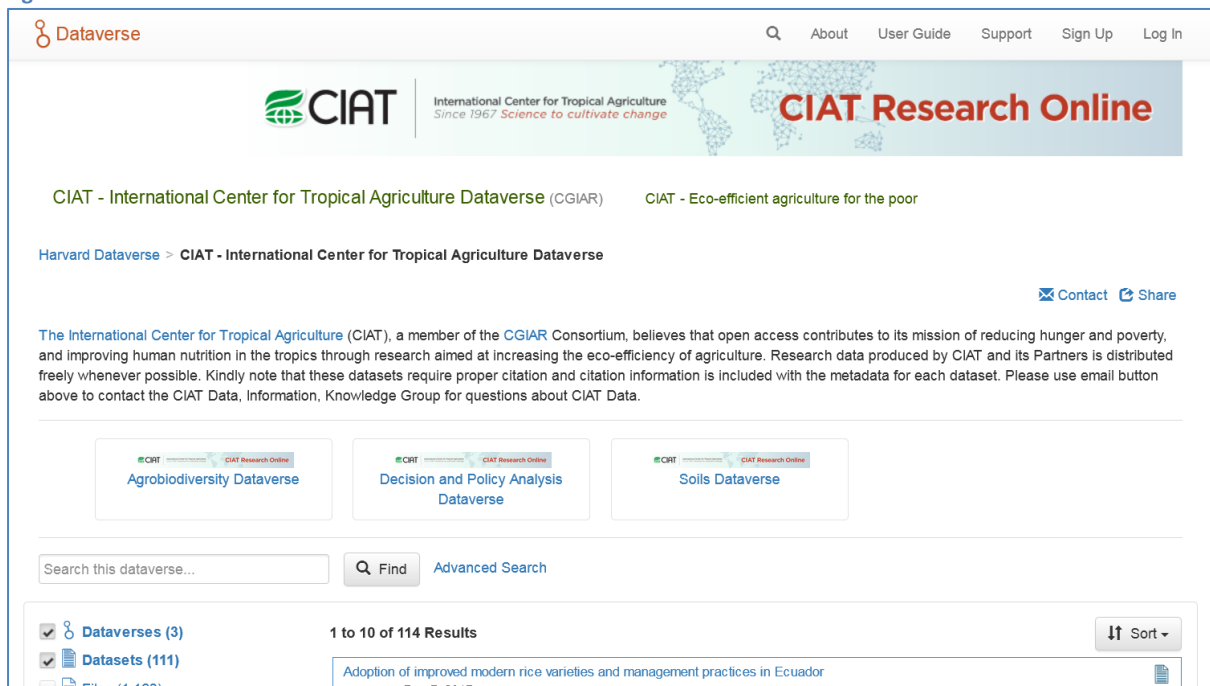
# Introduction to Dataverse

## Introduction

Dataverse is an open source web application for sharing, preserving, exploring and analysing research data. It was created by the Institute for Quantitative Social Science at Harvard University. Although aimed originally at social science data, a Dataverse can be created for research data on any subject. The website for the Dataverse Project is <https://dataverse.org/>

The image below shows the CIAT Dataverse which can be found at the following URL: <https://dataverse.harvard.edu/dataverse/CIAT>

Figure 1 - CIAT Dataverse



## Structure of a Dataverse

A Dataverse is effectively a container for “datasets” where each “dataset” comprises research data, code, documentation and metadata. A Dataverse can also contain other Dataverses, i.e. Dataverses can be nested. For example, in Figure 1 you can see that the CIAT Dataverse includes three nested Dataverses: Agrobiodiversity Dataverse, Decision and Policy Analysis Dataverse and Soils Dataverse.

## Citation

One thing that often stops researchers sharing their data is the belief that they won't get the credit for the work they have done. With Dataverse a unique citation is automatically generated when a dataset is created. For example, the unique citation for the CCAFS Household Baseline Survey in the CCAFS Dataverse is

CCAFS, 2015, "CCAFS Household Baseline Survey 2010-2012", [doi:10.7910/DVN/IUJQZV](https://doi.org/10.7910/DVN/IUJQZV), Harvard Dataverse, V2, UNF:6:h/b2JxlvusEKXLXeRTwC7Q==

For further information about data citation please see the Data Citation page on the Dataverse website at <http://best-practices.dataverse.org/data-citation/>

## Hosting

There are two options for hosting your Dataverse. The easiest is to have your Dataverse hosted at Harvard. This is a free service to all researchers. The infrastructure at Harvard is very good and great support is available to help you set up your Dataverse – you can also include your own branding in the header including working links so that your Dataverse has the look and feel of your own website. For example, the CIAT Dataverse includes a link in the header to take you directly to the CIAT website.

The other option is self-hosting and you can download and install the relevant software. This gives you more administrative control, but you would need an IT expert to install and manage the site including upgrading, taking backups, etc. You would also need good server infrastructure for hosting the application.

We would generally recommend the first option – i.e. having the Dataverse hosted at Harvard. To find out more about the two options see the online guides available on the Dataverse site. The User Guide is for those who wish to host their Dataverse at Harvard and the Installation Guide is for those who want to host their own Dataverse. The following link will take you to the Guides: <http://guides.dataverse.org/en/latest/>

## Permissions

When a dataset is released, the default is for public access. However, you can choose to restrict the entire dataset giving access to named users only. Alternatively, you can restrict individual files within a dataset even if the dataset itself has public access.

## Metadata

When you create a dataset, you will need to add metadata (formerly known as the Study Catalogue). The table below lists and describes the metadata elements for a dataset in Dataverse. Elements with a \* before the name are compulsory elements. We recommend you use this list as a template for your own datasets and work on completing this list throughout the project. From experience we can tell you that finding this information at the end of a project is very difficult as those who know are likely to have moved on to other activities.

<b><i>Element of Metadata</i></b>	<b>Description</b>
<i>*Title</i>	Full title by which the dataset is known
<i>Subtitle</i>	A secondary title used to amplify or state certain limitations on the main title
<i>Alternative title</i>	A title by which the work is commonly referred, or an abbreviation of the title
<i>Alternative URL</i>	A URL where the dataset can be viewed, such as a personal or project website
<i>Other ID</i>	Another unique identifier that identifies this dataset (e.g. producer's or another repository's number)
<i>*Author</i>	The person(s), corporate body(ies), or agency(ies) responsible for creating the work
<i>*Contact</i>	The contact(s) for this dataset

<b>Element of Metadata</b>	<b>Description</b>
<i>*Description</i>	A summary describing the purpose, nature, and scope of the dataset
<i>*Subject</i>	Domain-specific Subject Categories that are topically relevant to the Dataset. This is multiple response and you should select all relevant options from the list which is: <ul style="list-style-type: none"> <li>• Agricultural Sciences</li> <li>• Other</li> <li>• Engineering</li> <li>• Business and Management</li> <li>• Computer and Information Science</li> <li>• Earth and Environmental Sciences</li> <li>• Physics</li> <li>• Chemistry</li> <li>• Law</li> <li>• Medicine, Health and Life Sciences</li> <li>• Arts and Humanities</li> <li>• Social Sciences</li> <li>• Astronomy and Astrophysics</li> <li>• Mathematical Sciences</li> </ul>
<i>Keyword</i>	Key terms that describe important aspects of the Dataset
<i>Topic Classification</i>	The classification field indicates the broad important topic(s) and subjects that the data cover.
<i>Related Publication</i>	Publications that use the data from this dataset
<i>Notes</i>	Additional important information about the dataset
<i>Language</i>	Language of the dataset – select from the list
<i>Producer</i>	Person or organisation with the financial or administrative responsibility over this dataset
<i>Production Date</i>	Date when the data collection or other materials were produced (not distributed, published or archived)
<i>Production place</i>	The location where the data collection and any other related materials were produced
<i>Contributor</i>	The organisation or person responsible for either collecting, managing, or otherwise contributing in some form to the development of the resource
<i>Grant Information</i>	Grant information including grant agency, grant number, etc.
<i>Distributor</i>	The organisation designated by the author or producer to generate copies of the work including any necessary editions or revisions
<i>Distribution Date</i>	Date that the work was made available for distribution/presentation
<i>Depositor</i>	The person or the name of the organisation that deposited this dataset to the repository
<i>Deposit date</i>	Date that the dataset was deposited into the repository
<i>Time Period covered</i>	Time period to which the data refer. This item reflects the time period covered by the data, not the dates of coding or making documents machine-readable or the dates the data were collected. Also known as the span.
<i>Date of collection</i>	Contains the date(s) when the data were collected
<i>Kind of Data</i>	Type of data included in the file: survey data, census/enumeration data, aggregate data, clinical data, event/transaction data, program source code, machine-readable text, administrative records data, experimental



<i>Element of Metadata</i>	<b>Description</b>
<i>Series</i>	data, psychological test, textual data, coded textual, coded documents, time budget diaries, observation data/ratings, process-produced data or other
<i>Software</i>	Information about the dataset series
<i>Related Material</i>	Information about the software used to generate the dataset
<i>Related datasets</i>	Any material related to this dataset
<i>Other references</i>	Any datasets that are related to this dataset such as previous research on this subject
<i>Data sources</i>	Any references that would serve as background or supporting material to this dataset
<i>Origin of Sources</i>	List of books, articles, serials or machine-readable data files that served as the sources of the data collection
<i>Characteristic of Sources Noted</i>	For historical materials, information about the origin of the sources and the rules followed in establishing the sources should be specified
<i>Documentation and Access to Sources</i>	Assessment of characteristics and source material
	Level of documentation of the original sources

## Summary

Dataverse is primarily an archiving facility. However, you can create a Dataverse and start populating it early on in your project, gradually building the archive as each stage of the project is completed. Datasets are only made public once they are released, so you can continue building your archive until you are ready to release it. If you have a Data and Document store for your project, then creating your archive should be relatively straight-forward as you can keep the same structure.

## Associate Videos

Videos accompanying the original release of the CCAFS Data Management Support Pack in 2013 are available as a playlist on the Statistical Services Centre YouTube Channel at <https://www.youtube.com/channel/UCs7EU95YMjvhvNozJKCD92xQ/playlists>. These videos have not been updated since the original release but are mostly still relevant.

In particular the playlist includes the following videos about Dataverse:

**Introduction** **to** **Dataverse:**  
<https://www.youtube.com/watch?v=EGYuj1JM1Qc&index=12&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpj>

**Creating** **a** **Dataverse:**  
<https://www.youtube.com/watch?v=9dMtCvCpZNM&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpj&index=13>

**CCAFS** **Dataverse:**  
<https://www.youtube.com/watch?v=tr33h7TzFeY&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpj&index=14>

### 36. Introduction to DSpace (coming soon)

- *Data Management*
  - *Archiving*
- *Communications*
  - *Data dissemination/publicity*
- *Using the Data*
  - *Sharing and Access to data*
- *Reference*
- *When*
  - *Delivery of research products*
- *PI, Researcher*

### 37. Introduction to AgTrials (coming soon)

- *Data Management*
  - *Archiving*
- *Using the Data*
  - *Dashboards/data displays*
  - *Sharing and Access to data*
- *Reference*
- *When*
  - *Delivery of research products*
- *Researcher*

### 38. Introduction to CCAFS-Climate (coming soon)

- *Data Management*
  - *Archiving*
- *Technical*
  - *IT Systems (hardware, software, services)*
- *Reference*
- *When*
  - *Delivery of research products*
- *Technician*