INITIATIVE ON
Excellence in Agronomy

CGIAR

EiA IMPLEMENTATION GUIDE TO MEET DATA MANAGEMENT GOALS

Excellence in Agronomy Initiative of the CGIAR
December 2023

CONTENTS

What are EiA's aspirations, particularly those relating to data and analytics?

Overall objective of EiA: EiA aims to deliver an increase in productivity and quality per unit of input (agronomic gain) for millions of smallholder farming households in prioritized farming systems by 2030.

> *"The TRANSFORM Work Package deals with data & analytics to help achieve this objective."*

TRANSFORM's data and analytics outcome: At least 20 research and scaling partners use and share common, open and FAIR (findable, accessible, interoperable and reuseable) data, tools and analytics to support the co-creation of locally relevant agronomic solutions integrating climate-smart, inclusivity and sustainability dimensions and assessing their performance using standardized protocols.

The TRANSFORM Work Package houses technical experts to help Use Case teams and other EiA activities with data management, analytics and decision support. TRANSFORM has developed data management tools and analytical solutions to help achieve EiA goals – including commitment to BMGF regarding Open and FAIR data.

What are the expectations regarding EiA's Use Cases concerning data?

1. Sign the Data Sharing Agreement and adhere to its terms and condition

2. Use EiA questionnaires, surveys and protocols for key data collection stages (e.g., validation, piloting, MELIA) and activities (e.g., add-on survey), changing only what is essential to reflect the nature and focus of the Use Case.

3. Collect data that is FAIR[1] and make old datasets FAIR in conformance with accepted standards (also per CGIAR Open and FAIR Data Assets Policy[2]). There are several tools available to do this – open to all – and the default should be that these are used.

[1] Wilkinson, M.D., M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al. 2016. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

[2] https://cgspace.cgiar.org/bitstream/handle/10568/113623/CGIAR_OFDA_Policy_Approved_16April2021.pdf?sequence=1&isAllowed=y

4. Use ONA – EiA's data management system – OR – ensure that Use Case's data management systems provide data to TRANSFORM.
5. Make data open (also per CGIAR Open and FAIR Data Assets Policy): Upload cleaned, processed data in near real-time to EiA's agronomy database. **NOTE: If another database is in use, it must have an API and EiA must be able to pull data into the EiA database in near real-time.**
6. Help test tools, provide feedback on if and how they could be improved.
7. standardized protocols, trial designs (e.g., for MVP Follow validation).

Why must I live up to these expectations?

1. Our funders expect Open and FAIR data, available in near-real time.
2. We need to comply with CGIAR's public goods mandates, its Open and FAIR Data Assets Policy, and increasing calls for Open Science.
3. There is a moral imperative to find rapid, reliable solutions for pressing global agricultural challenges – and these rely on Open and FAIR data.
4. The need for high-quality Open and FAIR data, particularly from developing countries, is particularly urgent with the increasing interest in and applications of Artificial Intelligence in agriculture. Lack of data from appropriate agroecologies can result in analytical bias and inaccurate results.

> *Keeping data to yourself might allow you to publish a few papers for personal advancement – but publishing Open and FAIR data (as a data paper and in an open database) will enable analysis over large data aggregations and allow accelerated research and deeper, broader insights.*

What exactly is Open and FAIR data?

Open Data: Data assets can be freely used, reused (modified) and redistributed (shared) by anyone.

FAIR data: Data assets that are Findable, Accessible, Interoperable, and Reusable, as defined in Fig. 1.

Data that are fully findable and accessible (downloadable by all) are fully open...BUT not all open data are fully interoperable or reusable. EiA and BMGF stipulate data must be Open AND FAIR to accelerate research & innovation.

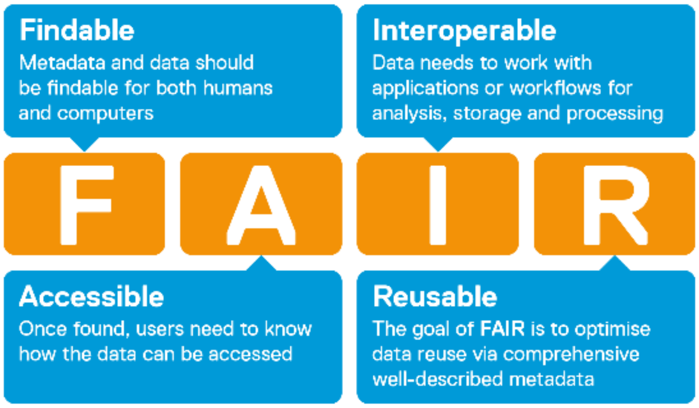

Findable
Metadata and data should be findable for both humans and computers
Interoperable
Data needs to work with applications or workflows for analysis, storage and processing
F
A
I
R
Accessible
Once found, users need to know how the data can be accessed
Reusable
The goal of FAIR is to optimise data reuse via comprehensive well-described metadata

Figure 1: What is FAIR data?

What tools can I use to share data that is Open and FAIR?

1. Develop and share ODK (Open Data Kit) questionnaires to collect standards-compliant (interoperable) data: **DataScribe** (https://datascribe.cgiar.org)
2. Aggregate data collected via ODK forms: **ONA** data management system (https://ona.io/home)
3. Develop fieldbooks to collect standards-compliant data from controlled (on-station) trials: **AgroFIMS** (https://agrofims.org)
4. Standardize « old data »: **FAIRscribe** (https://fairscribe.cgiar.org)
5. Standardize « old data »: **Carob** workflow (http://carob-data.org) – requires good knowledge of R
6. Provide access to standardized, FAIR data via an agronomy database: **Data Pool** (https://agronomydata.cgiar.org)
7. Archive data and metadata for longer term: **repositories** (e.g., EiA Dataverse)

For help with any of these tools, contact Céline Aubert (c.aubert@cgiar.org) or Medha Devare (m.devare@cgiar.org).

How do all these tools fit together?

EiA has a data ecosystem based on Open and FAIR data that builds on these and other tools. The visual below provides an overview of this ecosystem.

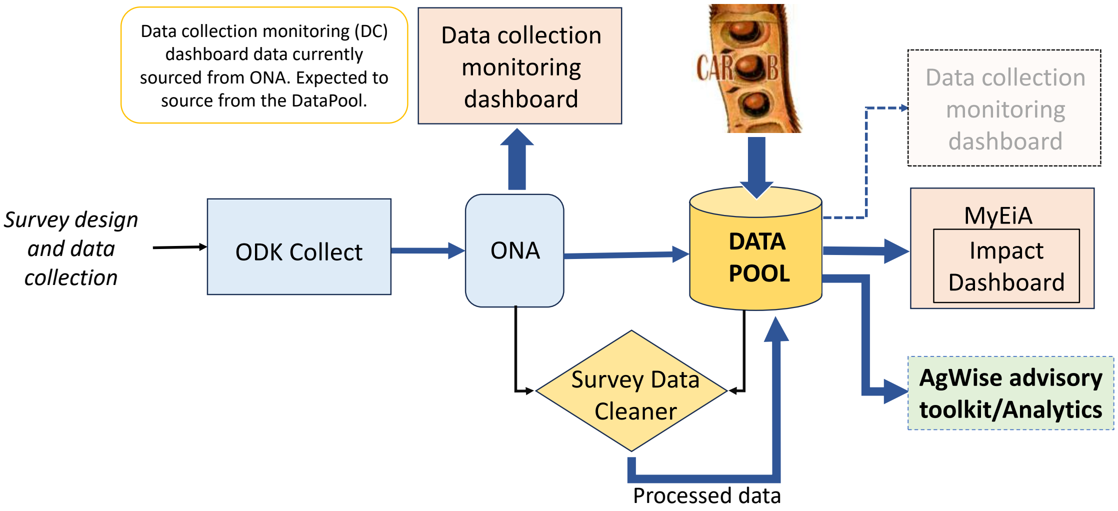

Data collection monitoring (DC) dashboard data currently sourced from ONA. Expected to source from the DataPool.
Data collection monitoring dashboard
CAROB
Data collection monitoring dashboard
Survey design and data collection
ODK Collect
ONA
DATA POOL
MyEiA
Impact Dashboard
Survey Data Cleaner
AgWise advisory toolkit/Analytics
Processed data

Figure 2: The EiA data ecosystem and tools.

How do I collect data that already adheres to EiA's agreed standards?

Depending on your data collection activity, you can use one of two existing tools to collect standards compliant data.

DataScribe (https://datascribe.cgiar.org)

DataScribe is a tool for building ODK questionnaires with standards-compliant variables, units and choice lists to make the data collected comparable, interoperable, interpretable, and easily aggregatable.

You should use this tool to develop and share ODK (Open Data Kit) questionnaires to collect community-agreed standards-compliant (and therefore interoperable) data.

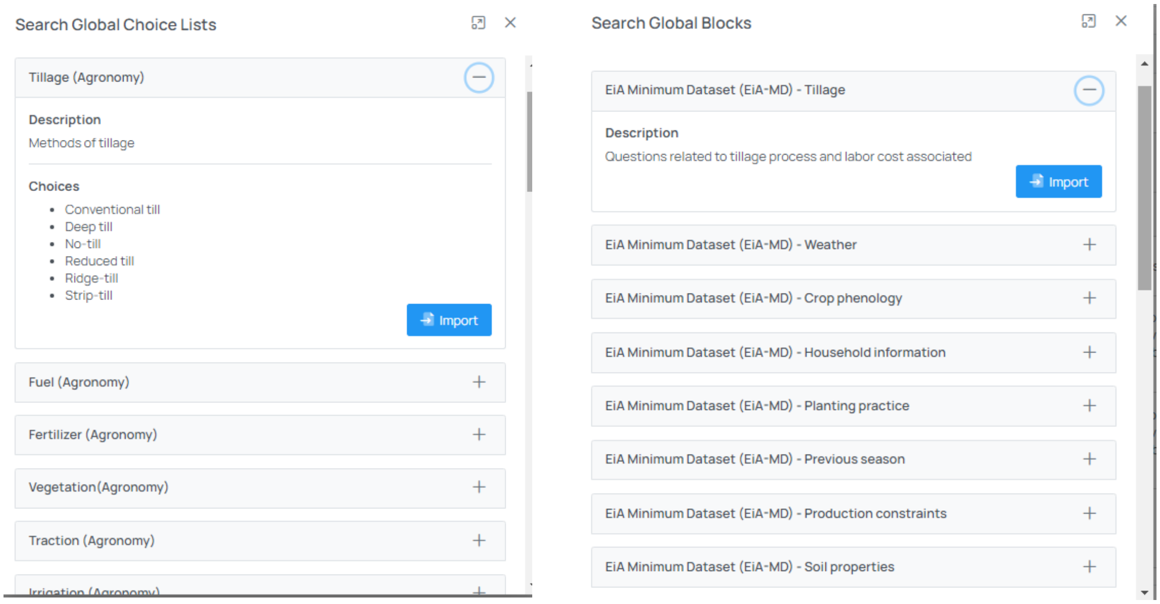
You will need an ORCiD (https://orcid.org) to use this and many other EiA tools. You can sign up for an ORCiD free of charge in just a few minutes. ORCID (Open Researcher and Contributor ID), is a global, not-for-profit organization allowing researchers to generate a unique, persistent identifier free of charge. This unique ID is an integral part of the wider digital infrastructure needed for researchers to share information on a global scale.

Once you log into DataScribe using your ORCiD, you will be able (1) upload an existing ODK form and add/edit/standardize questions or (2) create a new ODK form by creating a questionnaire and adding questions to it, standardizing key variables and units easily.

DataScribe offers:

- Lists of about 60 choice lists (850+ choices) with answer options relevant for agricultural surveys or experiments – enabling further standardization across surveys.
- 55 modules / blocks with questions related to particular topics, e.g., tillage.

- Blocks (groups) of questions from the RhoMIS tool (focused on socioeconomics) and EiA minimum data variables list – you can reuse entire groups of questions or select amongst these.



Search Global Choice Lists

Tillage (Agronomy)

Description

Methods of tillage

Choices

- Conventional till
- Deep till
- No-till
- Reduced till
- Ridge-till
- Strip-till

Import

Fuel (Agronomy)

Fertilizer (Agronomy)

Vegetation(Agronomy)

Traction (Agronomy)

Irrigation (Agronomy)

Search Global Blocks

EiA Minimum Dataset (EiA-MD) - Tillage

Description

Questions related to tillage process and labor cost associated

Import

EiA Minimum Dataset (EiA-MD) - Weather

EiA Minimum Dataset (EiA-MD) - Crop phenology

EiA Minimum Dataset (EiA-MD) - Household information

EiA Minimum Dataset (EiA-MD) - Planting practice

EiA Minimum Dataset (EiA-MD) - Previous season

EiA Minimum Dataset (EiA-MD) - Production constraints

EiA Minimum Dataset (EiA-MD) - Soil properties

Figure 3: Example of DataScribe choice lists and blocks.

For more information and details on how to use DataScribe, consult the user manual embedded in the tool. Help is also available on each screen of the tool.

If you don't use DataScribe or a tool that facilitates the collection of data adhering to these standards, your data:

- will not receive support from EiA teams to be analyzed nor be integrated in dashboards and other tools provided by EiA to support data management;
- will not be interoperable or easily understandable, and therefore difficult to use for analysis;
- will require substantial effort to render it reusable and of value beyond your immediate needs;
- is likely to be more error-prone, as digital data collection and the choice lists built into a tool like DataScribe minimize the possibility of errors by data collectors.

AgroFIMS (https://agrofims.org)

AgroFIMS is best-suited to collect standards-compliant data from controlled (on-station) trials. It allows users to easily design and create fieldbooks to collect agronomic data already tied to a metadata standard and standard variables, units etc. The tool offers an intuitive web-based platform to guide users through different screens following the operations called for in a typical agronomic experiment. It also enables the specification of brief protocols (how operations need to be conducted in the field). The tool stores agronomic study information in three major groups or modules not visible to the user (Figure 4).

The Research Management Information module allows the user to add metadata, including information about organizational and project affiliations, funding, personnel, sites, and crops associated with an experiment. The Study Design module allows formulation of an experiment or trial, and the Study Variables module is a group of standardized variables that include crop traits, soil fertilization, agronomic management practices, biotic and abiotic stressors, and soil and weather parameters. AgroFIMS aggregates this information in an easily usable data collection form or fieldbook. The fieldbook is loaded into a free Android application called KDSmart, allowing digital data collection.

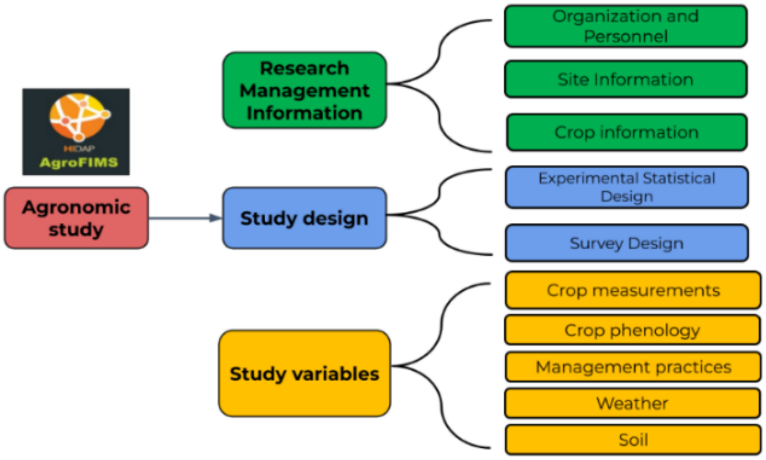


Figure 4. Indicative contents of the three main modules in AgroFIMS.

Guidance on AgroFIMS is available at: https://agrofims.org/documentation and more details in a paper (AgroFIMS: A tool to enable digital collection of standards-compliant FAIR data).

What should I do once I have used DataScribe to create an ODK form?

The ODK form created must be uploaded and saved to a folder in the ONA data management system (https://ona.io/home). EiA has an account that allows unlimited data submissions to ONA. Below are the brief steps to get you started on using ONA:

- You will need to first create a new account in ONA to use it.
- Then, share your ONA username with Céline Aubert (c.aubert@cgiar.org), EiA ONA account manager, to get access to a folder for each activity (Validation, Piloting, Add-on survey, ….).
- The Use Case lead will get admin rights on this ONA folder and be able to share it with colleagues and enumerators. The folder is accessible only by those people specified by the Use Case lead.
- The ODK form developed for an activity will be uploaded in the dedicated ONA folder.
- Data can now be collected using the ODK Collect Android app, which replaces more error-prone paper collection, supports DataScribe forms, and can work without network connectivity.
- Once the ODK Collect app is downloaded on your phone, a direct link between ODK and ONA is made by configuring ODK settings.
- When connected, the DataScribe form uploaded to ONA is accessible directly on your phone. For more detail, see https://help.ona.io/knowledge-base/guide-using-enketo-odk-collect/#odk-collect.

- Once enumerators/data collectors collect data in the field they must submit the ODK forms with the data to ONA, and the forms will appear in the folder.
- Any data collected at different timesteps using the same ODK form will be added to previously collected data. ONA therefore aggregates data from different timesteps in the same form.

EiA data scientists clean and process data using the ODK forms submitted to ONA for specific activities. This data is used to assess performance of Use Case-based recommendations and calculate Key Performance Indicators (KPIs).

It is therefore important to use standard ODK forms and to not change these, as any changes (1) make data processing and KPI calculations difficult and non-standard, and (2) create challenges in presenting a consistent visualization of results across EiA activities (via dashboards – see Figure x).

What should I do once I have created a fieldbook using AgroFIMS?

Once you have created a fieldbook in AgroFIMS, you can download it and export it to an Android device to digitally collect your data. To do this, you will need to download and install the KDSmart application on an Android device.

KDSmart information is available at: https://agrofims.github.io/helpdocs/collect/kdsmart/

The fieldbook should be saved and sent from the Android device to the primary investigator / project lead after each data collection stage. Once data collection is complete for all samples and times in your experimental season, the complete fieldbook can be exported from the Android device. The data is now ready for processing and upload to the agronomy database (see below).

How do I standardize legacy data?

FAIRscribe (https://fairscribe.cgiar.org)

FAIRscribe is a comprehensive web-based workflow to publish FAIR outputs (including datasets, publications, or any other digital resource), featuring:

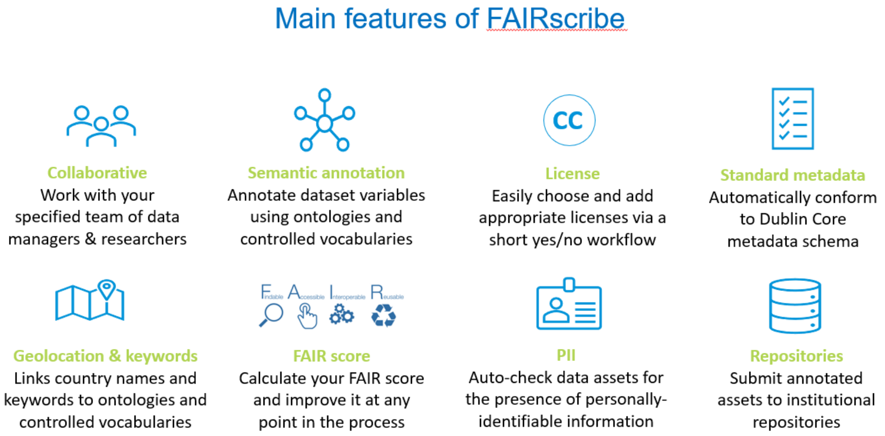
- Robust team management and collaboration features
- Flexible organization of resources
- Standards-driven metadata authoring with powerful automation features
- No-hassle publishing in any Dataverse or CKAN repository

FAIRscribe operates via a web interface (you need only to navigate web pages to use the tool) and – like most of EiA's tools – uses ORCiD authentication. It allows you / your data manager to standardize datasets and other assets (Fig. 5).

FAIRscribe makes it easy to:

- Organize your resources under multiple collections (e.g., as an EiA Use Case)
- Assign metadata authoring and review teams for each individual resource
- Collaboratively build multi-faceted, multi-lingual, standards-conformant metadata
- Check your FAIR score at any point and directly go to "recommendations" to improve it
- Easily specify the temporal dimensions of your dataset

- Specify spatial coverage using the UN M49 standard for regions and countries
- Credit contributors (people/organizations) using authoritative sources, among which are:
    - Individual authors
    - Institutional authors via the authoritative Research Organization Registry (ROR)
    - Curated funder lists from the ROR
- Identify the best suited license for your resource via an intuitive 5-step License Wizard - or attaching custom terms of use
- Automatically extract keywords for your resource, already mapped to standards
- Easily search and add keywords that are extracted from over 20 standard vocabularies
- Upload data files associated with the described resource, annotate tabular data at a column-level and incorporate your annotations in the dataset
- Request review of an annotated resource through the workflow prior to publishing to a desired Dataverse or CKAN repository



Main features of FAIRscribe

Collaborative
Work with your specified team of data managers & researchers

Semantic annotation
Annotate dataset variables using ontologies and controlled vocabularies

License
Easily choose and add appropriate licenses via a short yes/no workflow

Standard metadata
Automatically conform to Dublin Core metadata schema

Geolocation & keywords
Links country names and keywords to ontologies and controlled vocabularies

FAIR score
Calculate your FAIR score and improve it at any point in the process

PII
Auto-check data assets for the presence of personally-identifiable information

Repositories
Submit annotated assets to institutional repositories

Figure 5. Overview of the primary features of FAIRscribe

For more information and details on how to use FAIRscribe, consult the user manual embedded in the tool. Help is also available on each screen of the tool.

Carob (http://carob-data.org)

If you are an R coder (ideally at least intermediate level), you can contribute to the Carob project. Carob is an open-source, highly collaborative community project to which we encourage contributions for the benefit of global agricultural research and innovation. All data transformations are done with R scripts, making it easy to enhance the workflows as needs arise, and to correct mistakes.

Carob cleans and transforms agricultural research data from experiments and surveys into a standard format and aggregates individual data sets into larger data collections that can be used in further research. You can download compiled data from the Carob website or generate desired compilations yourself using the Carob scripts. See Fig. 6 for a view of the types of open data in the almost 600K records (as of November 2023) and their geographic spread.

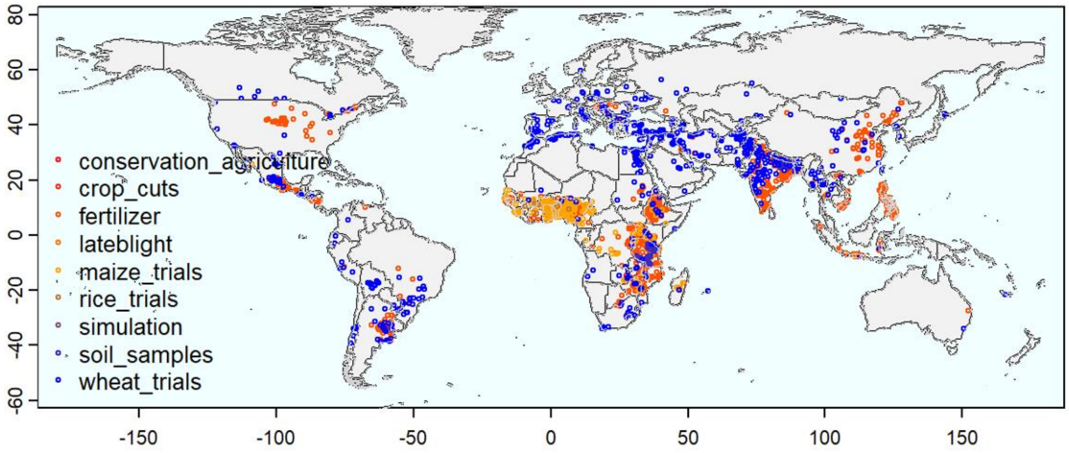


Figure 6. Types of data in the almost 600K Carob records (as of Nov 2023) and their geographic spread.

There is a substantial amount of raw primary research data available. This provides an opportunity to combine these data to address important additional research questions beyond the original purpose of one or a few datasets. However, it is very time consuming to reuse research data for some of the following reasons:

- Datasets have their own set of variable names, accepted values, and file structures. Even two files within a dataset may have discrepancies.
- Published data is often incomplete and needs to be augmented with information from publications.
- Most datasets have mistakes, especially in the location data and spelling. These mistakes can often be corrected (or removed) but doing that can be very time consuming.

Carob aims to solve these problems, and to provide a reproducible ET(A)L workflow (to extract, transform, aggregate and load data). Carob therefore makes it much easier to reuse raw research data. Once a script has been written to standardize a dataset, these data can be readily used by others as well. Or you can use a script and expand it, for example, to include additional variables, without having to start from scratch.

For more information and details on how to contribute to Carob, contact Eduardo Bendito (e.bendito@cgiar.org) or Robert Hijmans (rhijmans@ucdavis.edu).

How can I share or access data?

We in the agricultural sector can facilitate agronomic gain by increasing productivity, profitability and quality per unit of input for millions of smallholder farming households in prioritized farming systems. This is a key EiA goal, with the aspiration of achieving it by 2030 – but this goal can only be realized through rapid access and use of large amounts of high-quality data, particularly as data science and allied analytical tools (including AI) continue to advance at an unprecedented pace.
EiA's data tools come together in an ecosystem that culminates in access to standardized, FAIR data via an agronomy database or data pool as indicated in Fig. 2.

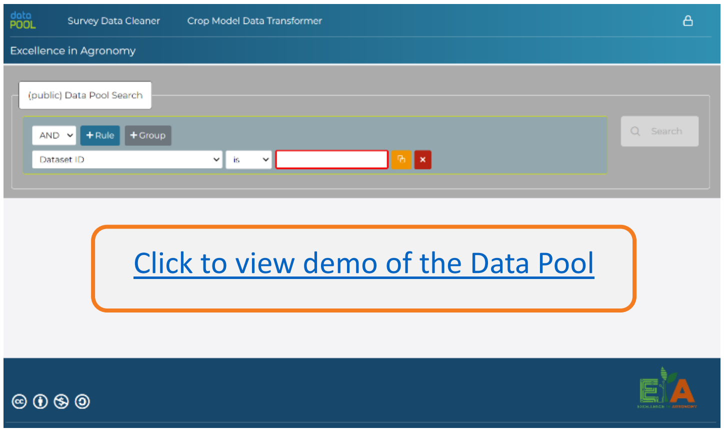
GARDIAN (https://gardian.cgiar.org)

GARDIAN is a metadata harvester and hub to find and access EiA's data-related tools. GARDIAN:

- Brings together data, publications, and interactive geospatial resources across 20+ repositories and sources.
- Enables access to over 219,000 publications and 26,000 datasets across CGIAR Centers and organizations working in the agricultural sector, presenting these in standard format, with the possibility of advanced search and filtering.
- 12+ TB of geospatial data on soils, 30+ crops, and future climate - visualized on a map interface and downloaded by desired bounding box or admin level.
- Offers all EiA tools under one umbrella – and more.

Data Pool (https://agronomydata.cgiar.org)

The Data Pool includes a private (authentication and permissions-based) database as well as a public database that use state of the art knowledge graph technology which relies on standards-compliant data as input. The Data Pool also further standardizes data as a service.



Click to view demo of the Data Pool

Figure 7. Click on the link in the image to see how the Data Pool works.

The Data pool allows users to:

- Search for data using terms that search within datasets (e.g., for maize data across Sub Saharan Africa with N applications of at least x kg/ha). See Fig. 7 for a short video demonstration.
- Develop downloadable databases flexibly for an infinite number of data products.
- Use the Crop Model Data Transformer to transform data from these searches to adhere to crop model (e.g., DSSAT) requirements. This eases data-to-model transformations that can otherwise take significant time.
- Download data that adheres to the Carob syntax, to ease the use of aggregated data in Machine Learning and other analytical applications.
- Clean ODK survey data (to remove long names, add codebooks) via the Data Pool's Survey Data Cleaner (https://agronomydata.cgiar.org/#/odkcl) prior to further analysis.
- Publish data immediately after collection (e.g., via AgroFIMS or ONA for DataScribe-based ODK data) to a private Data Pool, specifying permissions regarding who can access these data.

- Publish data directly to the public Data Pool or move it from the private Data Pool once data is clean and Personally-Identifiable Information has been removed (see below for what this entails) – **no more than 3 months after collection is completed.**

What constitutes PII and how do I deal with it?

PII, or Personally Identifiable Information, refers to any information that could enable the privacy of an individual or set of individuals to be violated. It includes a variety of information that must be dealt with before datasets are made public, as shown in Table 1.

Table 1. Exemplar PII types and how to deal with them.

Type of PII	How to deal with it
Names of individuals	○ Can upload to private Data Pool with PII; this Data Pool is generally more secure than keeping these data on your laptop ○ Must de-identify/re-code PII (e.g., replace names with numerical identifiers; 100 etc.) when uploading to public Data Pool
Names of households	○ Can upload to private (secure) Data Pool with PII ○ Must de-identify/re-code PII (e.g., replace names with numerical identifiers; HH101 etc.) when uploading to public Data Pool
Village names	○ Can upload to private (secure) Data Pool with PII ○ Must de-identify/re-code PII (e.g., replace names with numerical identifiers) when uploading to public Data Pool
Geocoordinates (e.g., households, villages)	○ Can upload to private (secure) Data Pool with PII ○ Obfuscate location coordinates by rounding to two decimal places (e.g., 43.47° N, 3.27° E instead of 43.4763° N, 3.2774° E) when uploading to public Data Pool

The Statistical Disclosure Control (SDC) Practice Guide provides excellent guidance on methods to deal with PII, including re-coding named variables and anonymizing location coordinates or other sensitive data. It also provides R scripts to make this easier. In general, no subjects should be surveyed without obtaining prior informed consent from the subjects, through a form approved by an Institutional Review Board/Institutional Review and Ethics Committee or similar entity. EiA has developed an Informed Consent template to make this easier; it is available on EiA SharePoint.

CGIAR

INITIATIVE ON
Excellence in Agronomy

Medha Devare, Chief Data Officer, *m.devare@cgiar.org*

Céline Aubert, Technical coordinator, *c.aubert@cgiar.org*

CGIAR is a global research partnership for a food-secure future. CGIAR science is dedicated to transforming food, land, and water systems in a climate crisis. Its research is carried out by 13 CGIAR Centers/Alliances in close collaboration with hundreds of partners, including national and regional research institutes, civil society organizations, academia, development organizations and the private sector. *www.cgiar.org*
We would like to thank all funders who support this research through their contributions to the CGIAR Trust Fund: www.cgiar.org/funders**.**
To learn more about this Initiative, please visit this webpage.
To learn more about this and other Initiatives in the CGIAR Research Portfolio, please visit www.cgiar.org/cgiar-portfolio

Citation: M.Devare and C.Aubert, 2023. EiA implementation guide to meet data management goals. Excellence in Agronomy Initiative, CGIAR, 14p.