CGIAR Data Harmonization Guidelines: Core Variables for FAIR and AI-Ready Data

December, 2025

Andrea Gardeazabal, Medha Devare, Kate Dreher, Marie-Angelique Laporte, Celine Aubert, Kai Sonder, Elizabeth Poole, Lorenzo Longobardi Carlo Azzarri, Satish Nagaraji

Version 1



Contents

1. Scope of this Document

2. How to Use this Guide

What this guide is (and is not)
When should you use this guide?
Step-by-step: how to use this guide in practice
Adding additional variables
What happens next?

3. Background and Context

4. Towards FAIR and AI Readiness

FAIR principles in CGIAR research data
From FAIR to AI-ready data
What does AI readiness mean in practice?
Relationship between AI readiness and FAIR

5. Core Variables and Guidelines

Purpose of core variables
Core variables as a minimum common denominator
A tiered approach to core variables
Domain-specific minimum sets
Cross-domain consistency
Use of standardized names, units, and protocols
Transparency and feasibility
Evolution of core variables

6. Moving Forward: Use, Feedback, and Evolution

7. Annex A. Domain-Specific Core Variables

A. Crop Breeding
B. Agronomy
C. Irrigation
D. Climate and Environment
E. Socioeconomics and Gender
F. Livestock

1. Scope of this Document

This document provides practical guidance to support CGIAR scientists and partners in collecting and managing research data in ways that enable interoperability, adherence to FAIR principles, and progressive AI readiness. It is written as a working guide for researchers, not as a policy or compliance document.

The scope of this guideline is intentionally focused and incremental. It introduces a minimum (core) set of variables for selected research domains and explains how these variables can be used as a shared baseline across CGIAR Centers. This first phase prioritizes feasibility, alignment, and transparency, recognizing the diversity of research contexts, disciplines, and operational constraints across CGIAR.

The guideline focuses on defining a common denominator for data collection rather than prescribing exhaustive variable lists. Core variables are intended to complement, not replace, project- or discipline-specific variables, and their use is meant to support comparability, reuse, and integration of data across projects and Centers.

This document covers the concept and purpose of core (minimum) variables, the initial domain-specific minimum variable sets for Agronomy, Breeding, Socioeconomics & Gender, Livestock, and Climate & Environment, and practical guidance on how to use the accompanying Excel files during the design of new data collection activities. It also provides recommendations on the use of standardized variable names, units, and protocols, as well as guidance on documenting context, constraints, and deviations where full alignment is not feasible.

At the same time, this guideline does not aim to define all variables required for any given research activity, replace scientific judgment or domain expertise, or mandate the retroactive restructuring of existing or legacy datasets. Detailed data management workflows across the full data lifecycle are outside the scope of this document, as is the definition of a final or fixed standard.

This document represents Phase 1 of a CGIAR-wide effort to harmonize data collection practices across Centers and domains. The core variables presented here reflect the minimum level of agreement reached at this stage and are expected to evolve over time through use, feedback, and collective discussion. Future phases will build on this foundation by refining variable sets, developing additional protocols, and strengthening cross-Center alignment.

The intended audience for this guideline includes CGIAR scientists, research teams, data managers, and technical staff involved in the design and implementation of data collection activities. The document is designed to be read together with the *How to use this guide* section, and to be used as a practical reference that supports consistency and transparency in data collection while preserving scientific flexibility and innovation.

2. How to use this Guide

This guideline is designed as a practical, step-by-step reference to support CGIAR scientists and partners in collecting data that is interoperable, FAIR, and progressively AI-ready. It is intentionally pragmatic and incremental.

This is not a fixed or final standard. It represents Phase 1 of a CGIAR-wide effort to agree on a minimum common set of data variables across research domains and Centers.

The guideline will evolve over time through consultation, testing, and collective agreement among the CGIAR Centers. The purpose of this guide is to help

researchers start from a shared baseline, while preserving scientific flexibility and contextual relevance.

What this guide is (and is not)

This guide is:

- A practical introduction to the concept of core (minimum) variables agreed at CGIAR level.
- A step toward greater data harmonization across Centers, projects, and domains.
- A reference for designing new data collection activities in a way that supports FAIR and AI-ready data.
- A shared starting point that will grow and improve over time.

This guide is not:

- A comprehensive list of all variables that should be collected.
- A rigid or prescriptive standard that replaces scientific judgment.
- A requirement to redesign existing datasets retroactively.
- A final or closed agreement.

The emphasis of this first phase is alignment, transparency, and feasibility, not perfection.

When should you use this guide?

You are encouraged to use this guide to assure that core data variables (at minimum) are present in datasets when:

- Designing new data collection activities (field trials, surveys, monitoring systems, experiments).
- Updating or revising data collection protocols.
- Preparing datasets for publication, sharing, or integration with other CGIAR data.
- Exploring opportunities to align data collection across projects or Centers.

For legacy datasets, this guide can be used as a reference where feasible, recognizing that full alignment may not always be possible. Legacy data alignment is not the primary objective of Phase 1.

Step-by-step: how to use this guide in practice

Step 1. Identify your research domain

This guideline provides domain-specific core variables for the following areas:

- Agronomy
- Breeding
- Socioeconomics & Gender
- Livestock
- Climate & Environment

Start by identifying the primary domain(s) relevant to your data collection activity.

Step 2. Review the core variables for your domain

For each domain, an Excel file is provided that lists the minimum (core) variables agreed as a common baseline across CGIAR.
These core variables represent the common denominator across Centers and projects. They are not intended to replace domain-specific or project-specific variables, but to complement them. Researchers are strongly encouraged to review these variables before launching data collection, to understand what is expected at minimum for interoperability and reuse.

Step 3. Use the Excel files during data collection design

The Excel files are intended to be used as a design-time tool, not only as a reporting template. When planning your data collection:

- Map your project variables against the core variables.
- Identify which core variables are already covered.
- Identify gaps and assess feasibility.

Step 4. Decide what can realistically be collected

It is recognized that not all core variables will be feasible in all contexts due to:

- Resource constraints
- Study design limitations
- Local conditions
- Ethical or operational considerations

In such cases:

- Prioritize transparency over completeness.
- Collect what is feasible, focusing on variables that help align different datasets (e.g., essential descriptors needed to interpret results across locations, time, or management conditions).
- Document constraints clearly in metadata and documentation.

Partial alignment is preferable to no alignment.

Step 5. Apply standard names, units, and protocols where possible

To support interoperability:

- Use the standardized variable names provided in the core variable lists.
- Convert locally collected units into the standardized units indicated in the Excel files before data publication or sharing.
- Follow recommended data collection protocols where indicated.

Standardized protocols improve data comparability and downstream reuse, including for AI-driven analysis.

Step 6. Document deviations and context

When deviations from the core variables or protocols occur, document:

- Why a variable was not collected or reported
- Why an alternative unit or method was used
- Any relevant contextual or methodological constraints

Clear documentation is essential for transparency, reuse, and trust in the data.

Adding additional variables

The inclusion of core variables does not limit the addition of other variables required by the research objectives.

Researchers are encouraged to:

- Collect additional variables relevant to their scientific questions.
- Maintain clear documentation distinguishing core variables from project-specific variables.

This approach supports both scientific innovation and cross-project interoperability.

What happens next?

This guideline and the associated core variable lists will be:

- Validated through use in real research contexts.
- Refined based on feedback from scientists across CGIAR.
- Expanded over time to include additional variables, protocols, and domains as agreed collectively.

By using this guide, researchers contribute directly to shaping the next phases of CGIAR's data harmonization efforts.

3. Background & Context

Data is one of CGIAR's most valuable scientific assets. Across Centers and programs, large volumes of data are generated every year through field trials, surveys, experiments, and monitoring activities. However, much of this data is still collected, documented, and stored in ways that limit its discoverability, interoperability, reuse, and long-term value.

In practice, datasets are often shaped by specific research questions, local practices, or project requirements, using different variable names, units, formats, and levels of documentation. While this diversity reflects the richness of CGIAR science, it also makes it difficult to integrate data across projects, Centers, and domains, and limits the potential for reuse beyond the original study.

At the same time, expectations around data management and sharing are increasing. Several of CGIAR's major funders have explicitly highlighted the importance of ensuring that the CGIAR data ecosystem is FAIR in practice, not only in principle. In practical terms, this means having data management processes that are not ad hoc, but implemented and operating across Centers, enabling scientific data to be accessed, integrated, used, and reused by different teams, programs, and partners over time.

Meeting these expectations requires more than making datasets available in repositories. FAIR data in practice depends on consistent variable definitions, standardized units, clear documentation, and shared conventions that allow data to move across projects and Centers with minimal additional effort. Without these foundations, access and reuse remain technically possible but operationally difficult.

The rapid development of artificial intelligence and advanced analytics adds a further dimension to this challenge. AI-driven methods rely on data that are not only FAIR, but also sufficiently complete, consistent, well-annotated, and free of systematic biases to support reliable analysis

and modeling. Data that are poorly documented, inconsistently structured, or ambiguous in meaning significantly reduce the effectiveness and trustworthiness of AI-based approaches.

Within CGIAR, the Digital Transformation Accelerator (DTA) is working with Centers and science programs to address these challenges in a practical and coordinated way. Rather than attempting to impose comprehensive standards across all research activities, the initial focus is on defining a small set of shared foundations that can be realistically adopted across diverse contexts.

This guideline represents one such foundation. It introduces a minimum (core) set of variables for selected research domains as a pragmatic starting point for improving data interoperability and readiness for advanced analytics. The emphasis is on identifying what is commonly needed across Centers, while allowing flexibility for project-specific and discipline-specific requirements.

By agreeing on and progressively adopting these minimum variable sets, CGIAR aims to move toward a data ecosystem in which datasets are easier to access, combine, compare, and reuse across projects and Centers. Over time, this approach is expected to reduce duplication of effort, improve returns on research investments, and enable more effective use of data for scientific discovery, decision support, and innovation.

Importantly, this guideline is designed to evolve. The core variables and recommendations presented here reflect an initial level of agreement and will be refined through use, feedback, and collective discussion among CGIAR Centers. The intention is not to define a final state, but to establish a shared baseline from which more advanced data harmonization and AI-ready practices can be developed incrementally.

4. Towards FAIR and AI Readiness

CGIAR has adopted the FAIR principles as a foundation for the management and stewardship of research data, with the aim of ensuring that scientific outputs are findable, accessible, interoperable, and reusable. The FAIR principles were formally articulated by Wilkinson et al. (2016) as a set of guiding concepts to improve the way scientific data and metadata are managed, shared, and reused.

FAIR does not prescribe specific technologies, platforms, or infrastructures. Instead, it defines characteristics that data and metadata should have in order to be reliably discovered, accessed, understood, and reused by both humans and machines. Central elements of FAIR include the use of persistent identifiers, rich and standardized metadata, shared vocabularies, and clear conditions for access and reuse. Importantly, FAIR does not imply that all data must be open; rather, data should be as open as possible and as closed as necessary, provided that access conditions are transparent and well documented.

Applied consistently, FAIR principles support transparency, reproducibility, and cumulative science. In practice, FAIR data enables researchers beyond the original study to locate relevant datasets, understand their content and context, integrate them with other data sources, and reuse them for new analyses. For an organization such as CGIAR, operating across multiple Centers, countries, and research domains, FAIR data is essential to enable collaboration and long-term scientific value.

However, emerging uses of data in advanced analytics and artificial intelligence require moving beyond basic FAIR compliance. AI-ready data builds on FAIR principles but places additional emphasis on specific data quality attributes that are critical for computational analysis. These include clarity and consistency in variable definitions, standardized units and coding schemes, completeness of key variables, structured annotation, and explicit documentation of assumptions, limitations, and potential sources of bias.

AI systems are particularly sensitive to ambiguity, inconsistency, and gaps in underlying data. When datasets are poorly documented or inconsistently structured, AI-based models may produce unreliable results or reinforce existing biases. For this reason, AI readiness should be

understood primarily as a data quality and governance challenge, rather than as a purely technical or modeling challenge.

Achieving greater AI readiness therefore starts at the point of data design and collection. Clear agreements on minimum data requirements, shared variable definitions, common units, and recommended protocols significantly improve the suitability of research data for both human interpretation and machine-based analysis. These practices also reduce the effort required for downstream integration, analysis, and reuse.

Within this context, the definition of minimum (core) variable sets represents a pragmatic step toward both FAIR implementation and AI readiness. By agreeing on a shared baseline across research domains, CGIAR creates the conditions for scientific data to be more easily accessed, integrated, and reused across projects and Centers, while preserving flexibility for project-specific and discipline-specific needs.

The approach outlined in this guideline is intentionally incremental. It does not assume immediate or universal use of AI methods, nor does it require all datasets to meet advanced AI-readiness criteria from the outset. Instead, it establishes a common foundation that improves data quality and interoperability today, and that can support more advanced analytical applications as capabilities, tools, and use cases evolve over time.

What does AI readiness mean in practice?

AI readiness refers to the extent to which research data can be reliably used for advanced computational analysis, including machine learning, statistical modeling, and decision-support systems. In the context of CGIAR science, AI readiness is less about the use of specific algorithms and more about the quality, structure, and clarity of the underlying data.

Several authors have emphasized that the performance and reliability of AI systems are primarily determined by data quality rather than model sophistication (Gebru et al., 2018; Suresh & Guttag, 2021). AI-ready data is therefore data that is sufficiently well defined, consistent, and documented to allow automated systems to interpret variables correctly and combine datasets across sources without extensive manual intervention. This requires clear and unambiguous variable definitions, consistent use of units and coding schemes, completeness of key fields, and explicit documentation of assumptions, limitations, and context.

Unlike human analysts, AI systems do not infer meaning from implicit knowledge or contextual familiarity. Ambiguities that researchers may resolve intuitively—such as inconsistent variable names, undocumented unit conversions, or locally defined categories—can lead to errors or misinterpretation when data are used in automated pipelines (Bender & Friedman, 2018). As datasets are increasingly reused beyond their original purpose, these issues become more pronounced.

AI readiness therefore begins at the point of data design and data collection. Decisions about which variables to collect, how they are defined, and how missing values, outliers, or exceptional cases are handled directly shape the downstream usability of the data. Once data are collected without sufficient structure or documentation, making them suitable for AI-driven analysis retrospectively becomes costly and, in some cases, infeasible (Batini et al., 2009).

Another critical aspect of AI readiness is transparency around data limitations and potential sources of bias. Research has shown that AI systems trained on incomplete, unbalanced, or systematically skewed datasets may amplify existing biases or produce results that are not robust across contexts (Gebru et al., 2018; Mehrabi et al., 2021). Clear documentation of sampling strategies, known constraints, and data gaps is therefore essential to support responsible and trustworthy use of AI-driven methods.

Importantly, AI readiness does not imply that all datasets must immediately support advanced machine learning applications. Rather, it implies that data are collected and documented in ways that preserve future analytical options. Even when AI is not an immediate objective,

improving AI readiness strengthens overall data quality, interoperability, and long-term reuse, aligning with principles of good scientific practice.

Relationship between AI readiness and FAIR

AI readiness builds directly on the FAIR principles. FAIR ensures that data can be found, accessed, interpreted, and reused by others (Wilkinson et al., 2016). AI readiness extends these principles by emphasizing internal data consistency, semantic clarity, and documentation quality required for reliable automated analysis.

FAIR is a necessary foundation for AI readiness, but it is not sufficient on its own. Data may be technically FAIR—discoverable and accessible—while still lacking the structure or consistency needed for AI-driven analysis at scale. As highlighted in recent discussions on FAIR for machine actionability, additional attention to data semantics, standardization, and quality is required to fully enable computational reuse (Mons et al., 2017; Wilkinson et al., 2019).

In this sense, FAIR enables data sharing, while AI readiness enables data actionability. Strengthening both together allows CGIAR to move toward a data ecosystem in which scientific data can be accessed, integrated, reused, and analyzed across projects and Centers, supporting both current research needs and future analytical capabilities.

The definition of minimum (core) variable sets presented in this guideline represents a concrete step toward this goal. By agreeing on shared foundations at the point of data collection, CGIAR supports FAIR implementation in practice while laying the groundwork for progressively more AI-ready data, without constraining scientific innovation or methodological diversity.

5. Core Variables and Guidelines

The core variables defined in this guideline represent a minimum common foundation for data that is collected and published across CGIAR research domains. Their purpose is to support interoperability, reuse, and progressive AI readiness, while preserving scientific flexibility and domain-specific rigor.

Importantly, core does not mean complete, nor does it imply that the same number or type of variables should apply across all domains. Differences in the size and composition of core variable sets reflect the intrinsic complexity, maturity, and data practices of each research domain, rather than differences in expectations or scientific standards.

The approach adopted here focuses on conceptual harmonization, rather than numerical uniformity.

Process for Defining the Core Variables

The core variables presented in this guideline were defined through a CGIAR-wide, participatory process coordinated under Area of Work 1 (Data Ecosystem) of the CGIAR Digital Transformation Accelerator. They are the result of collective discussion and agreement across Centers, rather than the work of a single institution or technical team.

The process was initiated during the Data Harmonization Workshop – Towards FAIR and AI-Ready Data, held from June 17 to June 20, 2025, as a hybrid event with in-person participation at CIMMYT headquarters in Mexico and remote participation from across the CGIAR system. The workshop brought together representatives from CGIAR Centers working across multiple scientific domains and data-related roles, with the explicit objective of agreeing on practical foundations for data harmonization.

Centers represented in the workshop included Alliance of Bioversity and CIAT, CIMMYT, IFPRI, and IITA through in-person participation, as well as AfricaRice, CIFOR-ICRAF, CIP, ICARDA,

ICRISAT, ILRI, IRRI, IWMI, and WorldFish through remote participation. In total, the workshop engaged participants from 12 CGIAR Centers.
During the workshop, participants worked collaboratively to review existing variable sets, identify common needs across Centers, and agree on a minimum shared baseline of core variables. Initial drafting focused on Agronomy and Socioeconomics & Gender as priority domains, using existing CGIAR initiatives as starting points. Following the workshop, multi-Center working groups were formed to refine these variable sets and extend the same approach to additional domains, including Crop Breeding, Livestock, Irrigation, and Climate & Environment.

The core variables included in this guideline therefore reflect a shared CGIAR agreement reached through structured discussion, practical testing, and cross-Center collaboration. They represent a first consolidated version of minimum core variables intended to support interoperability and reuse across CGIAR research, while allowing flexibility for domain-specific and project-specific needs.

Core variables as a minimum common denominator

Core variables are intended to function as a shared baseline that enables data to be understood, integrated, and reused across projects, Centers, and domains. They define the minimum level of information needed to make datasets interoperable and analytically meaningful beyond their original context.

Researchers are expected to continue collecting additional variables required by their specific scientific questions. Core variables are not a constraint on scientific design, but a mechanism to ensure that datasets retain value when combined with others or reused over time.

A tiered approach to core variables

To accommodate differences across research domains while maintaining a coherent framework, core variables are organized conceptually into two tiers. This tiered structure applies across all domains and is reflected in the accompanying Excel files, without requiring changes to their content.

Tier 1: Essential core variables

Tier 1 variables represent the minimum set of variables that are essential for basic interoperability and reuse within a domain. These variables are expected to be applicable to

most data collection activities within that domain and form the foundation for FAIR implementation in practice.

Tier 1 variables typically include:

- Variables required to identify the experimental unit or observation context
- Key outcome or response variables central to the domain
- Essential descriptors needed to interpret results across locations, time, or management conditions

Tier 1 variables define the lowest acceptable level of alignment for new data collection activities when interoperability and reuse are objectives.

Tier 2: Domain-critical core variables

Tier 2 variables represent **additional core variables that are critical for meaningful interpretation and integration of data within a domain**, but whose applicability may depend on study design, context, or resource availability.

These variables often capture:

- Domain-specific processes or conditions
- Contextual factors required for robust analysis
- Variables that substantially improve comparability and analytical depth when available

Tier 2 variables are strongly recommended where feasible, but their inclusion may vary depending on project scope and constraints. Together, Tier 1 and Tier 2 define the full core variable set for each domain.

Domain-specific minimum sets

The accompanying Excel files define the domain-specific core variables using a consistent structure, including standardized variable names, units, recommended protocols, and references to relevant ontologies where available.

While the number of core variables differs across domains, this reflects domain-specific realities:

- Crop Breeding and Irrigation domains tend to have relatively compact Tier 1 sets, reflecting highly standardized experimental designs and well-defined outcome variables.
- Agronomy includes a broader set of core variables to capture management practices and field-level heterogeneity.
- Climate & Environment and Socioeconomics & Gender require larger core sets due to the diversity of environmental, social, and contextual factors that influence interpretation.
- Livestock data inherently involves complex biological, management, and system-level variables, resulting in a larger core variable set.

These differences are expected and appropriate. Interoperability is achieved through consistent definitions, units, and documentation, not through equalizing the number of variables across domains.

Cross-domain consistency

Although core variables are defined by domain, a small set of cross-domain contextual information is implicitly required across all datasets, including clear identification of location, time, experimental or observational unit, and data collection context. These elements support integration across domains and are reflected in variable definitions and documentation practices, even where not explicitly listed as domain-specific core variables.

Use of standardized names, units, and protocols

To support FAIR implementation and AI readiness, researchers are strongly encouraged to use the standardized variable names and units provided in the core variable lists. Data may be collected using local units or conventions where necessary, but should be converted to standardized units prior to data publication or sharing.

Where recommended data collection protocols are provided, their use improves data comparability and downstream reuse. When deviations are necessary, they should be clearly documented.

Transparency and feasibility

It is recognized that full alignment with all core variables may not be feasible in all contexts. In such cases, transparency is essential. Researchers should document which core variables were not collected and the reasons for exclusion. Partial alignment, combined with clear documentation, is preferable to non-alignment.

Evolution of core variables

The core variable sets presented in this guideline reflect the current level of agreement across CGIAR Centers and research domains. They are intended as a shared starting point and are expected to evolve over time through practical use, feedback from researchers, and continued collective discussion across Centers.

As experience accumulates, future phases of this work may refine tier definitions, address redundancies, and strengthen cross-domain alignment, while maintaining flexibility for domain-specific needs and scientific innovation.

The domain-specific core variables referenced in this guideline are provided in a set of accompanying Excel files, included as annexes to this document. These Excel files constitute the authoritative and detailed representation of the core variable sets for each domain and should be used as the primary reference for variable specifications.

Each Excel file follows a consistent structure and includes, where available, standardized variable names, definitions, units, recommended data collection protocols, and references to relevant ontologies. The Excel files are designed to serve as practical working tools during the design of data collection activities and the preparation of datasets for sharing and reuse.

The tiered presentation of variables described in this guideline (Tier 1 and Tier 2) is intended to support clarity and implementation. It does not alter the content of the Excel files, nor does it introduce additional requirements. Researchers should always refer to the Excel annexes as the authoritative source for technical details and variable definitions.

By adopting this incremental and tiered approach, CGIAR establishes a shared foundation for interoperable, FAIR, and progressively AI-ready data, while respecting the diversity of research domains and preserving scientific autonomy across Centers.

6. Moving forward: use, feedback, and evolution

This guideline represents a shared starting point for improving data interoperability, FAIR implementation in practice, and progressive AI readiness across CGIAR research domains. Its value will ultimately be determined not by its completeness, but by its use in real research contexts.

Researchers are encouraged to apply this guideline pragmatically when designing new data collection activities, updating protocols, and preparing datasets for sharing and reuse. Partial alignment, combined with clear documentation of constraints and context, is preferable to non-alignment. Transparency and feasibility are core principles of this first phase.

The minimum (core) variable sets and tiered structure presented here reflect the current level of agreement across CGIAR Centers and domains. They are expected to evolve as experience accumulates, as redundancies are identified, and as opportunities for stronger cross-domain

alignment emerge. Feedback from scientists, data managers, and partners using these guidelines in practice will be essential to inform future refinements.

Future phases of this work will build on the foundations established here by refining core variable definitions, strengthening protocol guidance, improving cross-domain consistency, and supporting the integration of legacy data where feasible. These efforts will continue to be developed through collective discussion and agreement across CGIAR.

By using and contributing to this guideline, researchers support a shared data ecosystem in which scientific data can be more easily accessed, integrated, reused, and analyzed across projects and Centers—strengthening the long-term value, impact, and credibility of CGIAR science.

References

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys, 41*(3), 1–52. https://doi.org/10.1145/1541880.1541883

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6*, 587–604. https://doi.org/10.1162/tacl_a_00041

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92. https://doi.org/10.1145/3458723

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35. https://doi.org/10.1145/3457607

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use, 37*(1), 49–56. https://doi.org/10.3233/ISU-170824

Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Proceedings of the 2021 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*, 1–9. https://doi.org/10.1145/3465416.3483305

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. O. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, Article 160018. https://doi.org/10.1038/sdata.2016.18

Wilkinson, M. D., Sansone, S.-A., Schultes, E., Doorn, P., Bonino da Silva Santos, L. O., & Dumontier, M. (2019). A design framework and exemplar metrics for FAIRness. *Scientific Data, 6*, Article 174. https://doi.org/10.1038/s41597-019-0184-5

7. Annex A. Domain-specific core variables (Tier 1 and Tier 2)

This annex provides a structured presentation of the domain-specific core variables defined in the accompanying Excel files. To improve clarity and usability, variables are presented using a tiered structure. This tiering is intended as a communication and implementation aid and does not replace or modify the authoritative variable inventories contained in the Excel annexes.

Tier 1 variables represent the essential baseline required to enable basic interoperability, reuse, and interpretation of data within a domain. Tier 2 variables represent domain-critical variables that substantially improve analytical depth and comparability but may be context-dependent or constrained by resources.

The presence and size of Tier 2 vary across domains, reflecting intrinsic differences in domain complexity and data practices.

A. Crop Breeding

Tier 1 – Essential core variables

Latitude; Longitude; Crop; Germplasm name; Planting date; Irrigated ; Gender; Age; Nitrogen fertilizer; Phosphorus fertilizer; Potassium fertilizer.

B. Agronomy

Tier 1 – Essential core variables

Administrative unit level 0 (country); Administrative unit level 1; Administrative unit level 2; Latitude; Longitude; Crop; Variety; Planting date; Harvest date; Yield (dry); Irrigated (Y/N); Gender; Age; Soil texture; Nitrogen fertilizer; Phosphorus fertilizer; Potassium fertilizer; Maturity date; Yield (fresh); Crop product.

C. Irrigation

Tier 1 – Essential core variables

Irrigation event date; Irrigation method; Flow rate of irrigation technology; Plot area; Water source; Crop type under irrigation; Irrigation start time; Irrigation end time; Soil moisture status before irrigation; Soil moisture status after irrigation; Energy source for pumping; Basis for irrigation scheduling.

D. Climate & Environment

Tier 1 – Essential core variables

Air temperature; Precipitation; Wind speed; Solar radiation; Soil moisture; Discharge; Groundwater level; Terrain elevation; Optical vegetation index (NDVI); Land cover class code.

Tier 2 – Domain-critical variables (recommended where feasible)

Water temperature; Dissolved oxygen; Biological oxygen demand; Chemical oxygen demand; Electrical conductivity; Total oxidised nitrogen; Nitrate in groundwater; Orthophosphate; pH; Pathogen indicator (E. coli); Reservoir level; Environmental flow; Irrigated crop cultivated area; Irrigated crop harvested area; Rainfed crop cultivated area; Rainfed crop harvested area; Water withdrawal (irrigation, livestock, domestic, energy, mining, manufacturing); Water supply coverage (piped); Physical water loss in drinking water systems; Irrigation water-use efficiency; Proximity to roads; Proximity to settlements; Irrigation-related time series variables where applicable.

E. Socioeconomics & Gender

Tier 1 – Essential core variables

Administration unit level 0 (country); Administrative unit level 1; Administrative unit level 2; Latitude; Longitude; Interview date; Age; Sex; Years of education; Primary occupation; Access to land; Total land area; Crop; Crop area; Crop production; Crop use; Farmgate price; Access to livestock; Livestock type; Income from labor; Income from business; Income from public sources; Other income; Household Dietary Diversity Score (HDDS); Minimum Dietary Diversity for Women (MDD-W); Access to extension services; Access to credit.

Tier 2 – Domain-critical variables (recommended where feasible)

Contact information; Relationship to household head; Months of residence; Education grade; Marital status; Ethnicity; Kitchen garden; Parcel-level variables (parcel number, status, use, size, ownership, water regime, distance, decision-making, control); Labor time variables by gender and activity; Cost variables (seed/variety, fertilizer, crop protection, labor, mechanization, livestock); Livestock production and use variables; Additional income sources; Agricultural assets; Social network indicators; Time to market.

F. Livestock

Given the diversity of livestock production systems, core variables are organized using a module-based approach.

Tier 1 – Cross-cutting baseline variables

Production system type; Biosecurity characteristics (as applicable); Disease prevalence indicators (as applicable); Mortality by age/sex and year type; Product or animal price; Feed input costs; Labor type (family/hired) and basic labor input; Access to credit or financial services.

Tier 2 – System-specific modules (recommended where applicable)

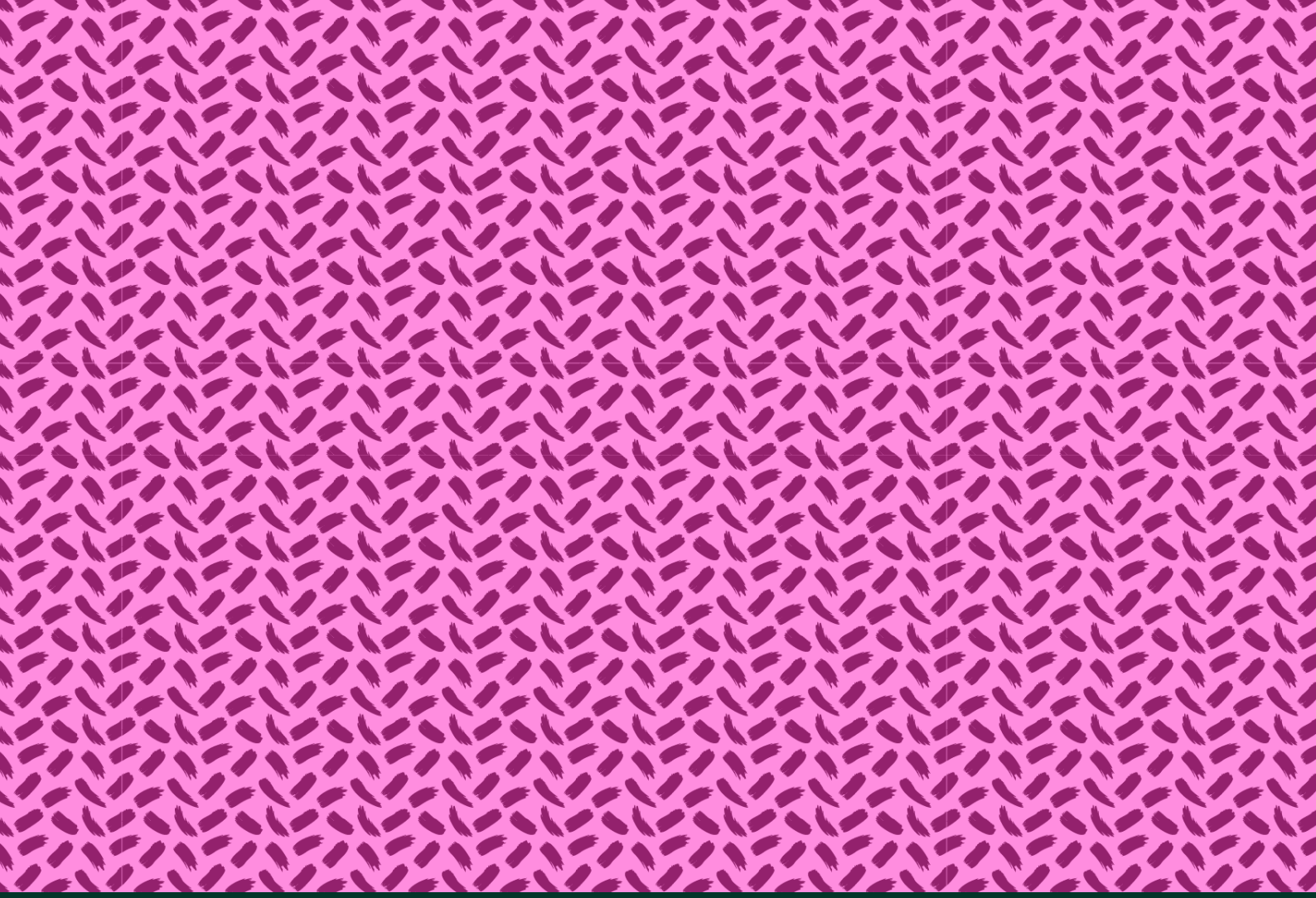
Tier 2 variables are defined within system-specific modules provided in the Excel annexes, including but not limited to:

- Small-scale fisheries
- Aquaculture
- Large ruminants
- Small ruminants
- Pig breeding and pig fattening
- Backyard poultry
- Poultry (layers and broilers)

These modules include additional variables related to reproduction, feeding, health, management practices, productivity, and economics that are required for robust analysis within each production system.

Note on use and evolution

This tiered presentation reflects the current level of agreement across CGIAR Centers and domains and is intended to support practical implementation. The tier definitions and domain groupings are expected to evolve in future phases as experience accumulates, redundancies are addressed, and cross-domain alignment is strengthened.



CGIAR

BETTER DIETS AND NUTRITION | BREEDING FOR TOMORROW | CAPACITY SHARING | CLIMATE ACTION | DIGITAL TRANSFORMATION | FOOD FRONTIERS AND SECURITY | GENDER EQUALITY AND INCLUSION

GENEBANKS | MULTIFUNCTIONAL LANDSCAPES | POLICY INNOVATIONS | SCALING FOR IMPACT | SUSTAINABLE ANIMAL AND AQUATIC FOODS | SUSTAINABLE FARMING