Data Collaborative Report
Session 1

October. 2025

Gardeazabal Monsalve, A.; Devare, M.; Dreher, K. A.; Koo, J.; Garcia Lopez, D. E.; Laporte, M. A.; Juarez, H.; Ramirez, D.; Attaher, S.; Domelevo Entfellner, J.-B.; De Leon, D.; Mercado, E. F.; Reyes, M.; Rinza, J.; Burac, M. A.; Gulles, A.; Mwanzia, L.; Garcia Andarcia, M.; Fonteyne, S. G.; Amagnide, A.; De Boeck, B.; Al-Sham'aa, K.; Raman, A.; Imoro, S.; Poole, E. J.; Sallan, M. A.; Jimenez, D.; Sonder, K.; Gakhar, S.; Ng, E. H.; Dhulipala, R.; Radanielson, A.; Govind, A.; Ali, I.; Azzarri, C.; Rayco, M. A.; Bernardo, M.; Ramos, J.; Kollegala, R.; Bhosale, S.

Version 1

1. Background and Objectives

The Data Collaborative is CGIAR's first Center of Excellence under the Digital Transformation Accelerator (DTA), established to strengthen data governance, interoperability, and reuse across Centers, science programs, and domains. The Collaborative responds to a shared recognition that while CGIAR generates a large volume of high-quality scientific data, inconsistencies in variables, metadata, and practices limit interoperability, FAIRness, and AI-readiness.

Session 1 marked the formal launch of the Data Collaborative and had three main objectives: (i) to introduce and validate the purpose, scope, and governance structure of the Data Collaborative; (ii) to review and collect feedback on the proposed Terms of Reference (ToRs); and (iii) to initiate collective alignment on a first set of core data variables, starting with agronomy and crop breeding as the most mature domains.

---

2. Session Overview

Date: October 28, 2025
Format: Virtual (Teams + Miro)
Duration: 90 minutes
Participants: Representatives from CGIAR Centers (CIMMYT, IFPRI, Alliance Bioversity-CIAT, CIP, ICARDA, ILRI, IRRI, IWMI, AfricaRice, ICRISAT-IN, IITA) and domain experts in agronomy, crop breeding, data management, and digital transformation.

The session combined short presentations with interactive exercises using a Miro board to gather real-time feedback. The emphasis was on co-creation, transparency, and learning rather than decision finalization.

---

3. Introduction to the Data Collaborative

The session opened with a detailed introduction to the Data Collaborative, positioning it as a practical and coordinated mechanism to advance data governance, interoperability, and reuse across CGIAR. Rather than proposing a new centralized system or repository, the Data Collaborative was framed as a shared governance and validation space, where Centers and science programs collectively agree on how data should be produced, described, and published in order to be reusable across contexts.

A key emphasis was placed on the shift from data integration to data interoperability. Participants were reminded that while integration efforts often require significant technical investment and alignment on platforms, interoperability focuses on ensuring that data produced in different places can meaningfully work together. This includes alignment on variables, metadata, ontologies, and protocols that allow datasets to be combined, compared, and reused without reprocessing them from scratch.

The domain-based approach was presented as a pragmatic entry point into this broader challenge. By organizing data discussions around domains such as agronomy, crop breeding, livestock, climate and environment, and socioeconomics,

the Collaborative aims to address commonalities across Centers while respecting disciplinary specificities. Agronomy and crop breeding were selected as starting domains due to their relative maturity and the existence of prior harmonization work that could be built upon.

The Data Collaborative was also clearly positioned within CGIAR's existing governance architecture. Rather than operating as a standalone initiative, it is intended to connect directly with existing technical, ethical, legal, and leadership structures, ensuring coherence and legitimacy of decisions.

---

4. Terms of Reference: Summary and Validation

The Terms of Reference (ToRs) were presented as the foundational document defining the mandate, scope, and operating model of the Data Collaborative. Participants were invited to view the ToRs not as a fixed or final instrument, but as a living framework that would evolve as the Collaborative matures.

The ToRs outline a multi-layered governance structure designed to balance agility with accountability. At the core is the Convening Team, led by the Digital Transformation Accelerator, which acts as the operational secretariat responsible for coordination, facilitation, and follow-up. This team works closely with the Core Working Group, composed of nominated Center representatives, who serve as the primary validation and decision-making body for cross-CGIAR data-related agreements.

Additional layers of governance provide safeguards and external perspective. The Technical Advisory Group is consulted selectively when proposals raise legal, ethical, cybersecurity, or risk-related considerations, ensuring alignment with institutional policies without slowing routine work. External Advisors, drawn from leading global organizations, provide strategic guidance and benchmarking against international best practices.

During the session, participants provided substantive feedback on the ToRs. Much of this feedback focused on clarifying roles and responsibilities, particularly for Center representatives, and on making decision-making processes more explicit. There was also interest in defining concrete indicators of success, such as adoption of agreed standards or measurable improvements in FAIR and AI-ready datasets. Overall, the feedback reflected strong endorsement of the Collaborative's purpose, with suggestions aimed at strengthening clarity, accountability, and implementation.

---

5. Core Variables: Framing and Rationale

A key objective of the Data Collaborative is to initiate collective alignment on a first set of core data variables, starting with agronomy and crop breeding. This was also addressed during the session, starting with clarification on what is meant by *core variables* before diving into specific domains. Core variables are not intended to replace rich, domain-specific datasets. Instead, they represent the minimum set of variables that should be consistently published to enable interoperability and reuse across Centers and domains.

Key framing points included: - Core variables define a baseline, not a ceiling - Extended variables will continue to exist and may be standardized later - The focus is on data publication, not all stages of the data lifecycle - Initial emphasis is on future datasets, with legacy data addressed later

---

6. Agronomy Core Variables: Discussion and Feedback

The agronomy segment formed a substantial part of the session and served as a first practical test of the Data Collaborative's working approach. The proposed agronomy core variables were introduced as a minimal yet sufficient set of variables that would allow agronomic datasets to be understood, aggregated, and reused across Centers.

The rationale for the proposed variables was grounded in previous large-scale harmonization efforts, particularly experiences derived from standardizing legacy datasets. These experiences demonstrated that many datasets become unusable for secondary analysis due to missing or inconsistent basic information such as location, crop establishment, management practices, or harvest outcomes.

Participant feedback was extensive and constructive. Several contributions highlighted the need to better capture experimental context, including elements of experimental design, treatment information at plot level, and distinctions between on-station and on-farm trials. Others emphasized the importance of variables related to soil health, marketable yield, or system classification (e.g., rainfed versus irrigated), particularly from a farmer-relevance and investment perspective.

At the same time, participants raised important questions about keeping the core set truly minimal. In particular, the relevance of demographic variables such as age and gender was debated, with broad agreement that their inclusion should be conditional on the type of study. There were also calls for clearer guidance on measurement units, moisture adjustment for yield, and machine-readable standards.

Overall, the discussion reinforced the value of the core-variable concept while highlighting the need for careful boundary-setting between what is essential for interoperability and what belongs in extended, use-case-specific datasets.

[Placeholder for Miro Board Image – Agronomy Core Variables Feedback]

---

7. Crop Breeding Core Variables: Discussion and Feedback

The crop breeding discussion followed the same structure as agronomy but reflected the distinct characteristics of breeding data and workflows. The proposed core variables were intentionally fewer, with a strong emphasis on alignment with agronomy wherever possible to support cross-domain interoperability.

The discussion acknowledged the diversity of breeding activities, ranging from early-stage nurseries to advanced multi-location trials. Participants emphasized that not all experiments generate the same types of data, particularly with respect to yield, phenology, or plot structure. This led to a nuanced discussion about the need to

distinguish between different types of breeding activities when defining expectations for core variables.

Feedback underscored the importance of spatial information at plot level, including row and column identifiers, to enable spatial analysis and correct modeling of field variability. There was also strong interest in improving the clarity and consistency of germplasm-related variables, ensuring that breeding lines, checks, landraces, and released varieties can be handled coherently across datasets and systems.

Participants also highlighted opportunities to align breeding core variables with existing standards and systems, such as the Breeding API and enterprise breeding platforms, to facilitate automated data exchange and reduce reporting burden. At the same time, there was agreement that variables related to farmer demographics are generally not relevant for most breeding trials and should not be considered core.

The discussion reinforced the importance of balancing interoperability with flexibility, particularly in a domain as methodologically diverse as crop breeding.

[Placeholder for Miro Board Image – Crop Breeding Core Variables Feedback]

---

8. Cross-Cutting Insights

Several cross-cutting themes emerged across both domains: - Strong support for the concept of core variables as an interoperability baseline - Recognition that extended variables and domain-specific depth remain essential - The need to harmonize overlapping variables across domains (e.g., location, germplasm) - Interest in leveraging existing systems (e.g., Enterprise Breeding System) to operationalize standards - Clear expectation that governance decisions translate into practical implementation

---

9. Next Steps

The following next steps were agreed or implied:

1. Consolidate and synthesize feedback from the Miro boards and post-session inputs
2. Revise agronomy and crop breeding core variable sets accordingly
3. Finalize and endorse the Data Collaborative Terms of Reference
4. Prepare guidance documents for using and publishing core variables
5. Extend the process to additional domains (livestock, climate & environment, socioeconomics)
6. Convene the next Data Collaborative session (scheduled for November 28)

---

10. Conclusion

Session 1 successfully launched the Data Collaborative as a shared governance and learning space for data harmonization across CGIAR. The session demonstrated

strong engagement, constructive debate, and a shared commitment to making CGIAR data more interoperable, FAIR, and AI-ready. The feedback collected provides a solid foundation for refining both governance mechanisms and technical standards in the next phase of work.

Annex A. Approved Terms of Reference for the Data Collaborative

A.1 Overview

The Data Collaborative is CGIAR's first Center of Excellence under the Digital Transformation Accelerator (DTA). Its purpose is to establish a coordinated, federated, and system-wide approach to data governance, interoperability, and reuse across CGIAR Centers and programs. The Collaborative aims to ensure that data produced across the system are FAIR-aligned, AI-ready, and fit for reuse across scientific domains and institutional priorities.

The Data Collaborative functions as a validation and alignment mechanism rather than as a data repository or delivery platform. Its role is to review, agree on, and promote shared standards, variables, protocols, and governance practices that enable interoperability while respecting Center autonomy and existing systems.

A.2 Governance Structure

Convening Team (Core Leads)
The Data Collaborative is coordinated by the Digital Transformation Accelerator Director, who serves as Lead Convenor. The Convening Team acts as the secretariat of the Collaborative and is responsible for planning, coordination, facilitation of meetings, documentation of decisions, and monitoring of progress.

The Convening Team includes representatives from the System Organization Digital Transformation Team and from the Data Ecosystem, Enabling Environment, Action Lab, and Digital Futures Areas of Work under the Accelerator. This structure ensures operational continuity and alignment with CGIAR system priorities.

Center Representatives (Core Working Group)
Each CGIAR Center nominates a focal point with relevant expertise in science, data, IT, policy, or operations. Together with the Convening Team, these representatives form the Core Working Group, which serves as the primary validation and decision-making body of the Data Collaborative.

Center Representatives are responsible for contributing domain expertise, validating proposed standards and frameworks, and facilitating internal coordination within their Centers to support implementation of agreed decisions.

Technical Advisory Group (TAG)
The Technical Advisory Group provides internal oversight on legal, ethical, cybersecurity, and risk-related aspects of data governance. The TAG is consulted on an ad hoc basis when proposals have implications for compliance or institutional risk. It does not operate as a standing committee, allowing for targeted engagement while maintaining quality assurance.

Members include representatives from legal, ethics and business conduct, IT and cybersecurity, risk management, and the CGIAR Integrated Corporate Initiative.

External Advisors
External Advisors are senior experts from global organizations who provide strategic guidance and external perspective. They are engaged once or twice per year and do not participate in routine decision-making. Their role is to help ensure alignment with international best practices and emerging trends in data governance and digital public goods.

A.3 Roles and Responsibilities

The Data Collaborative is responsible for:

- Reviewing and agreeing on core data variables and ontologies across scientific domains
- Validating protocols for data collection, cleaning, storage, and publication
- Promoting FAIR and AI-ready data practices across CGIAR
- Coordinating capacity building and knowledge sharing related to data governance
- Ensuring alignment with ethical, legal, and cybersecurity requirements

Proposals validated by the Data Collaborative may be escalated to the Global Science Team (GST) and, when formal endorsement is required, to the Global Leadership Team (GLT).

A.4 Operating Model

The Data Collaborative meets virtually on a quarterly basis, with additional ad hoc meetings convened as needed to address specific topics. One in-person meeting is planned each year to review progress, consolidate learning, and define priorities for the following cycle.

Decisions are reached through consensus wherever possible. When consensus cannot be achieved, the Convening Team documents differing perspectives and proposes a way forward, which may include escalation to the GST or GLT.

A.5 Initial Priorities and KPIs

Initial priorities of the Data Collaborative include the validation of core variables for priority domains, the development of guidance for FAIR and AI-ready data publication, and the establishment of clear implementation pathways across Centers.

Indicative KPIs include the number of Centers adopting agreed standards, the number of datasets published using validated core variables, and measurable improvements in data interoperability and reuse across programs.

A.6 Closing Note

The Terms of Reference establish the Data Collaborative as a practical, inclusive, and scalable mechanism for strengthening data governance across CGIAR. By providing a shared foundation for alignment and decision-making, the Collaborative

enables more consistent, transparent, and impactful use of data to support science, innovation, and institutional learning.

Annex B. Feedback Collected During Data Collaborative Session 1 (Verbatim)

This annex compiles the feedback collected during the first Data Collaborative session, captured live on the Miro board and through facilitated discussion. The feedback is presented verbatim and organized by thematic area for transparency and traceability. It reflects participants' perspectives at the time of the session and informed subsequent refinements of both the Terms of Reference and the proposed core variables.

B.1 Feedback on Data Collaborative Terms of Reference

Group composition and representation

- Core Group composition: Some Core Group roles may have changed (e.g., Senior Manager of Digital Transformation or Digital Core).
- Group composition: Include a Research Data Management leader who is familiar with data protection laws.
- Representation: Is each Center expected to have one Core Group representative? Can we get a current list of representatives?
- Multiple representatives: Can a Center have more than one representative?
- Center structure: How will each Center be organized—one overall lead, then leads per domain?

Roles, responsibilities, and decision-making

- Core Group responsibilities: Add a clear description of representatives' responsibilities—such as ensuring that agreed decisions are implemented within their Centers.
- Decision-making: How will decisions be made within the group? Include a short section on this process in the document.
- KPIs: KPIs are mentioned but not defined. Which specific metrics will be used to measure success (e.g., number of Centers adopting shared standards, datasets made FAIR/AI-ready, etc.)?
- Governance coordination: Clarify how the Data Collaborative's governance bodies (Convening Team, Center Representatives, Technical Advisory Group, External Advisors) will work together to make CGIAR's data ecosystem FAIR, AI-ready, and aligned across Centers.

B.2 Feedback on Agronomy Core Variables

What to add

- For field experiments, include experimental design and factors; for surveys, specify the sampling method.

- Add “Farm size in hectares (total area of land cultivated by the household).”
- Include a categorical variable “Type of system” (Rainfed or Irrigated).
- Add region and/or sub-region grouping for countries.
- For field experiments, include row and column information at the plot level.
- Add “Crop market segments,” which typically span across borders.
- Include “Soil organic carbon.”
- Add “Marketable yield,” particularly relevant for root and tuber crops.
- Add “Tillage.”
- Add total harvested/planted area (important for investment prioritization).
- Add information related to treatment in the specific plot to better understand the context of the values.

What to remove

- “Gender” and “Age” are not valid for in-station or plot-level agronomy experiments. Even for farmers’ plots, these parameters are often not relevant.

What to change

Variable naming - Suggest changing variety_name to germplasm_name, as testing may involve materials that are not commercially released varieties. - Use variety_name or germplasm_name to cover all plant materials used in agronomic trials, and include an additional field variety_status (e.g., released, experimental, landrace) where relevant. - Harmonize agronomy and breeding datasets, particularly for location and variety/germplasm fields. - Use the scientific name instead of the common name for crops.

Location and geospatial information - Consider using a bounding box instead of only longitude/latitude to map plots more accurately. - For household surveys, latitude and longitude should refer to the field, not the farmer’s home. - Clarify whether latitude/longitude refers to the field center point or its boundary.

Measurement standards and units - Report yield at a standard moisture content (e.g., 14%). - Report yield in kg/ha instead of t/ha and ensure units are machine-readable (e.g., kg_per_ha).

Demographic data - Age and gender are swapped. - Gender and age may be relevant for on-farm or demonstration trials and should be specified accordingly.

Data recording level - Clarify that variables should be recorded at the plot level, not the trial level.

B.3 Feedback on Crop Breeding Core Variables

What to add

- Include plot identity using row and column numbers to enable spatial modeling.
- Include yield measurements; fertilizer and water parameters alone are insufficient for breeding analysis.

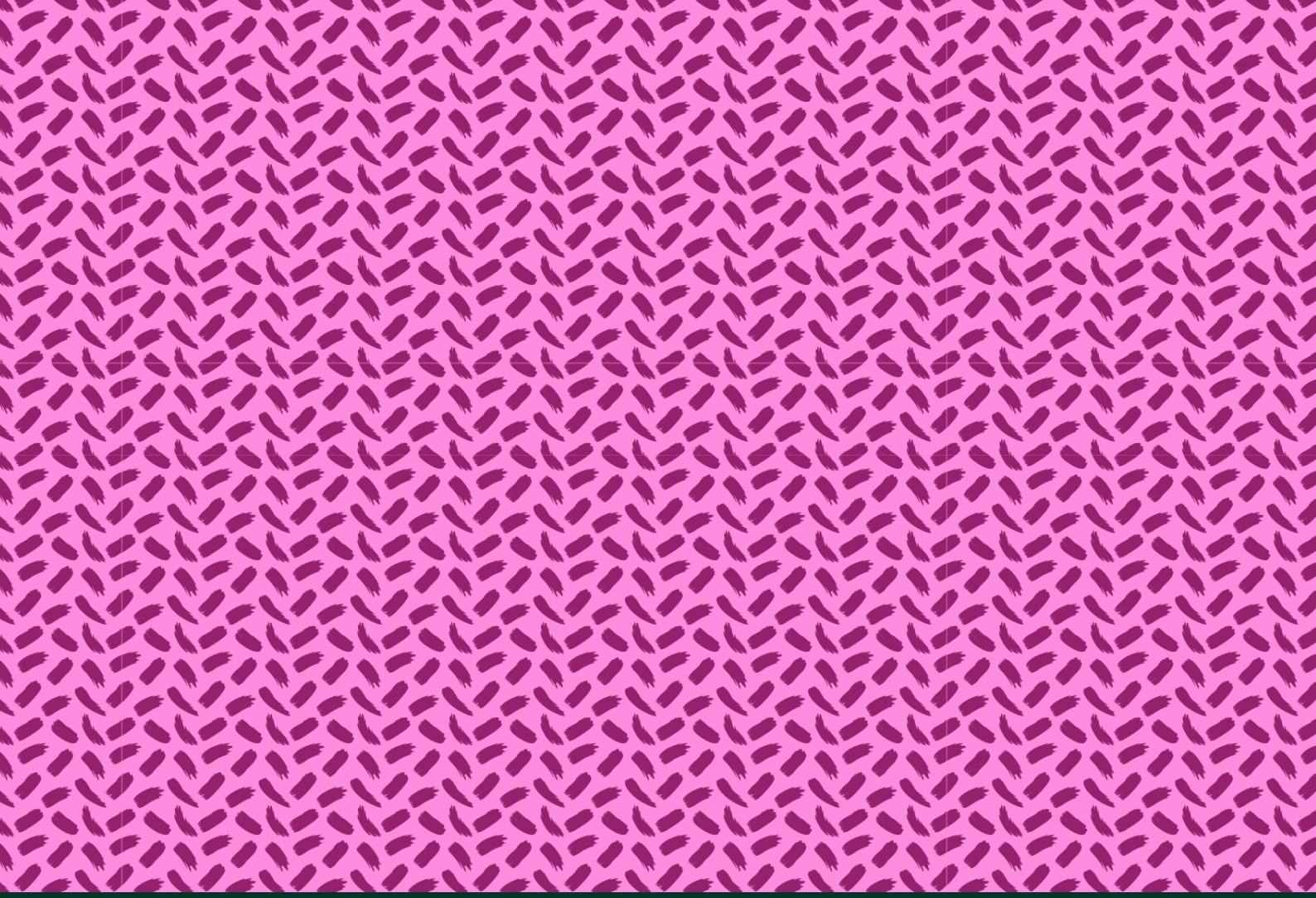
- Include plot yield (in grams) and plot size.
- Add type of trial (e.g., nursery, advanced yield trial).
- Include experimental design, breeding stage, breeding pipeline, experiment year, season, and planting coordinates.
- Add germplasm-related fields such as role in trial (e.g., check), crossing year, and target traits.
- Include breeding program name, plot size, and key phenotypic traits.
- Record initial and final plant stand counts as covariates.
- Include administrative levels to mitigate geolocation errors.
- Add maturity-related variables (flowering time, maturity time, harvest date).
- Specify the observational unit (plant, plot, etc.).
- Include visual scoring for abiotic stress as a covariate.
- Include germplasm ID for provenance.
- Add “Ecosystem” and “Cropping season.”
- For trials, include plot yield in kg/ha (adjusted for moisture content).

What to remove

- Dry yield is not always evaluated in early-stage breeding experiments.

What to change

- Harmonize breeding variables with agronomy where possible.
- Use the standard germplasm_name from the Enterprise Breeding System (EBS) with a categorical data type.
- Consider using bounding boxes instead of only longitude/latitude.
- Map variables to the Breeding API (BrAPI) ontology to ensure interoperability.
- Gender and age are not relevant for most breeding trials and should be excluded from core variables.
- Use variety_name / germplasm_name consistently to encompass all plant materials used in breeding trials.



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