TECHNICAL REPORT

FAIRification of CIP Legacy Datasets

Henry Juarez[a], Vilma Hualla[a], Piero Palacios[a], and Leroy Mwanzia[a]

[a] *International Potato Center, Lima Peru*

INFO

Submitted	31 December 2025
Keywords management,	*FAIR data* Core Sets of Variables, Agricultural research datasets, AI-ready data CIP Dataverse
Area of Work	*Digital Futures*
HLO	*3.1.1*

EXECUTIVE SUMMARY

The International Potato Center (CIP) has undertaken a structured FAIRification process to improve the quality, interoperability, and reuse of its legacy and ongoing research datasets hosted in CIP's Dataverse. Aligned with CGIAR FAIR and AI-ready data principles, this effort focuses on the adoption of Core Sets of Variables validated at the CGIAR system level, embedding standardized data practices across key research domains. As an initial priority, CIP has FAIRified legacy datasets in crop breeding, agronomy, plant health (late blight), and in-situ crop diversity, cleaning and harmonizing existing records and assigning persistent DOIs to ensure traceability and long-term access. In total, dozens of high-value legacy datasets—comprising hundreds of field books and multiple years of observations—have been standardized and made interoperable for downstream analyses, meta-studies, decision support systems, and AI-enabled applications. In parallel, CIP proposes a minimum Core Set of Variables for biodiversity to address the specific requirements of in-situ crop diversity datasets, complementing existing agronomy and breeding standards. For plant health and disease epidemiology, the FAIRification process builds on the agronomy core while incorporating domain-specific extensions suitable for late blight monitoring, modeling, and decision support. Looking forward, CIP commits to embedding the Core Set of Variables by default in all future breeding datasets from 2026 onward, ensuring FAIR compliance at the data collection and management stages rather than through retrospective curation. This FAIR-by-design approach strengthens data quality, reduces future curation costs, and positions CIP's data assets as AI-ready, reusable resources that directly support CGIAR's system-wide data governance and digital transformation objectives.

1. Problem Statement

In alignment with CGIAR FAIR and AI-ready data principles, the International Potato Center (CIP) has initiated a structured process to ensure that its legacy and ongoing datasets comply with Core Sets of Variables validated at the CGIAR system level.

As an initial focus, CIP has implemented Core Sets of Variables for agronomy and crop breeding datasets, ensuring consistency in variable definitions, metadata, and documentation to support interoperability, reuse, and AI readiness. These standards are being applied to legacy datasets hosted in CIP's Dataverse and are embedded by default in new datasets moving forward.

In parallel, CIP will progressively extend this approach to additional data domains, including plant health, social sciences, and biodiversity, to support the preparation of AI-ready datasets aligned with CGIAR system-wide data governance and ecosystem objectives.

To address datasets related to in-situ crop diversity and agrobiodiversity, CIP additionally proposes a minimum Core Set of Variables for biodiversity, covering geographic, taxonomic, and contextual descriptors. This proposed core set is intended to complement existing agronomy and breeding standards and to support harmonization, cross-site comparison, and integration with environmental and socio-ecological data.

2. Legacy datasets prioritized

The following legacy datasets have been prioritized for FAIRification and publication in CIP's Dataverse due to their strategic value, scientific relevance, and potential for reuse across CGIAR programs:

1. In-situ Crop Diversity (legacy)
   - Puno ZABD (Peru), Cusco hotspot (Peru), and Cumbal hotspot (Colombia)
2. Peru Potato Breeding Trials (legacy)
   - LTLB and LBHT breeding cycles, B1 and B3 population
3. Late Blight Trials (legacy)
   - Late blight trials (legacy)

3. In-situ Crop Diversity (Legacy): Core Variables for Biodiversity

The In-situ Crop Diversity legacy datasets from Puno ZABD (Peru), Cusco hotspot (Peru), and Cumbal hotspot (Colombia) represent a distinct data domain that requires a biodiversity-oriented Core Set of Variables, complementary to the Core Sets currently applied to agronomy and crop breeding datasets.

Existing datasets

The following in-situ crop diversity datasets are currently available in CIP's Dataverse:

- Puno ZABD (Peru). This dataset includes molecular marker data from ullucus, oca, mashua, and native potato accessions collected in Cuyocuyo (Puno, Peru) and genotyped using DArTSeq and Illumina SNP array platforms. The data correspond to collections conducted between 2023 and 2025 and support analyses of genetic diversity and population structure of Andean crops conserved in situ within the Andenes de Cuyocuyo Agrobiodiversity Zone.
    - https://doi.org/10.21223/P3/QPHCEQ
    - https://doi.org/10.21223/P3/M613LS
    - https://doi.org/10.21223/P3/TFOJNZ
    - https://doi.org/10.21223/P3/FYYLLO
- Cusco hotspot (Peru). The historical native potato database covers records from 1927–2020 and was expanded in 2022 to 15,257 observations, including surveys of 335 farming families. The study was conducted in Challabamba, Colquepata, and Paucartambo (Cusco, Peru), a high-altitude and agroecologically diverse area suitable for native potato cultivation.
    - https://doi.org/10.21223/P3/UXWTSU
- Cumbal hotspot (Colombia). In Cumbal (Nariño, Colombia), native potato agrobiodiversity is conserved in situ within traditional shagra systems managed by indigenous families. Between 2023 and 2025, citizen science and survey-based monitoring, supported by VarScout, strengthened documentation, local knowledge, and community participation.
    - https://doi.org/10.21223/P3/T0GRN9
    - https://doi.org/10.21223/PD5PPN
    - https://doi.org/10.21223/GIRW3W

Proposed minimum Core Set of Variables for Biodiversity

To support harmonization, interoperability, and reuse across sites, the following variables are proposed as a minimum Core Set of Variables for biodiversity, to be consistently defined across all in-situ crop diversity datasets.

Geographic and administrative variables

- country
- adm1 (first administrative level)
- adm2 (second administrative level)
- latitude
- longitude

Biological and taxonomic variables

- crop
- species
- variety (or landrace name)

Recommended additional minimal variables

To enhance biodiversity relevance and analytical value while maintaining a lightweight structure, the following additional variables are recommended:

- collection_year — temporal context for diversity assessments
- altitude_m — key for agrobiodiversity and climate gradient analyses
- source_type (e.g., farm household, community seed system, wild/feral)
- local_name — supports linkage to local and traditional knowledge
- use_category (e.g., food, processing, cultural)

The definition and validation of this Core Set of Variables for Biodiversity through the CGIAR Data Collaborative would support consistency across Centers and enable integration with agronomic, environmental, socio-ecological, and climate datasets, contributing to the preparation of FAIR and AI-ready biodiversity data.

4. Potato Breeding Trials

For Potato Breeding Trials, CIP has undertaken a FAIRification effort focused on the systematic cleaning, harmonization, and standardization of datasets already published in CIP's Dataverse, with full adoption of the Core Set of Variables for Crop Breeding.

Current status

- Thirty-eight (38) breeding datasets hosted in CIP's Dataverse have been cleaned and FAIRified.
- These datasets collectively contain approximately 200 field books, representing multiple breeding cycles and multi-location evaluations conducted in Peru.

The curated datasets cover:

- Breeding cycles for late blight resistance (LTLB)
- Advanced heat-tolerant LBHT clones
- B1 and B3 breeding populations across different recombination cycles
- Participatory varietal selection (PVS) trials evaluating resistance to late blight and frost across diverse agroecological regions
- Parental value trials for tuber yield under high-temperature conditions

The FAIRified breeding datasets are publicly available in CIP's Dataverse under the following DOI links:

- https://doi.org/10.21223/P3/EKOYZ1
- https://doi.org/10.21223/P3/LDMFOM
- https://doi.org/10.21223/P3/CINPSO
- https://doi.org/10.21223/P3/RWIMFO
- https://doi.org/10.21223/P3/QXIM59
- https://doi.org/10.21223/P3/SKYDRT
- https://doi.org/10.21223/P3/KZWGT8
- https://doi.org/10.21223/P3/2WCUAE
- https://doi.org/10.21223/P3/M40FYT
- https://doi.org/10.21223/P3/4FTDO8
- https://doi.org/10.21223/P3/BOXOEZ
- https://doi.org/10.21223/P3/SFXXDC
- https://doi.org/10.21223/P3/TNNRYW
- https://doi.org/10.21223/P3/VYCLVX
- https://doi.org/10.21223/P3/1GET14
- https://doi.org/10.21223/P3/UGWAZL
- https://doi.org/10.21223/P3/9VMENB
- https://doi.org/10.21223/P3/MWOJGR
- https://doi.org/10.21223/P3/OOQ73N
- https://doi.org/10.21223/P3/H50YAO
- https://doi.org/10.21223/P3/IO6NAV
- https://doi.org/10.21223/P3/NQBNWX
- https://doi.org/10.21223/P3/DCQB18
- https://doi.org/10.21223/P3/F8IGI9
- https://doi.org/10.21223/P3/XGKXGE
- https://doi.org/10.21223/P3/YWXFWM
- https://doi.org/10.21223/P3/3WAPNU
- https://doi.org/10.21223/P3/CDCKNH
- https://doi.org/10.21223/P3/QJ10B7
- https://doi.org/10.21223/P3/KATPZ8
- https://doi.org/10.21223/P3/RVCHKV
- https://doi.org/10.21223/P3/MO4PSJ
- https://doi.org/10.21223/P3/LCGD4P
- https://doi.org/10.21223/P3/OTYRIV
- https://doi.org/10.21223/P3/XVRHST
- https://doi.org/10.21223/P3/FFORMJ
- https://doi.org/10.21223/RHSVIY
- https://doi.org/10.21223/7WC7FM

Forward-looking commitment

- All future breeding datasets (from 2026 onward) will include the Core Set of Variables by default, embedding FAIR compliance directly into data collection and management workflows rather than applying it retrospectively.
- This approach institutionalizes FAIR practices within CIP's breeding programs and significantly reduces future

curation and harmonization costs, while improving data quality, interoperability, and reuse potential from the outset.

5. Late Blight Trials (Legacy)

The Late Blight legacy datasets constitute a specialized plant health and disease epidemiology domain that, while building upon the Core Set of Variables for Agronomy, requires an additional minimum Core Set of Variables specific to plant pathology to adequately support disease monitoring, epidemiological modeling, and decision support applications.

For these datasets, CIP has adopted the framework and variable definitions proposed by the CGIAR Data Collaborative: Center of Excellence on Data, using the Core Set of Variables for Agronomy as a common baseline and extending it with plant pathology–specific variables where required.

Current status

- Thirty late blight epidemics, together with associated yield data, have been uploaded to CIP's Dataverse following the Core Set of Variables for Agronomy and the proposed plant pathology extensions.
- Each dataset has been assigned a persistent DOI, ensuring traceability, citability, and long-term access.

The FAIRified late blight datasets are available in CIP's Dataverse at the following DOI links:

- https://doi.10.21223/P3/KFTL9L
- https://doi.10.21223/P3/QUYFEF
- https://doi.10.21223/P3/IWQD1Q
- https://doi.10.21223/P3/GUXJCB
- https://doi.10.21223/P3/YU81LE
- https://doi.10.21223/P3/1T7MCI
- https://doi.10.21223/P3/D8ZV2N
- https://doi.10.21223/P3/W2JTGQ
- https://doi.10.21223/P3/BAOBQL
- https://doi.10.21223/P3/6JNTVM
- https://doi.10.21223/P3/3HYQ86
- https://doi.10.21223/P3/NFZPMZ
- https://doi.10.21223/P3/GTV7Y4
- https://doi.10.21223/P3/UZVDW6
- https://doi.10.21223/P3/XD4APA
- https://doi.10.21223/P3/XKKVJS
- https://doi.10.21223/P3/XIF8Q9
- https://doi.10.21223/P3/DNJKC3
- https://doi.10.21223/P3/YPJ8H1
- https://doi.10.21223/P3/WKLS5U
- https://doi.10.21223/P3/0F9T62

Core Agronomy Variables Relevant to Plant Disease

The following variables from the Core Set of Variables for Agronomy are considered essential for plant disease and DSS datasets, as they provide the agronomic, spatial, and temporal context required for disease assessment, interpretation, and modeling.

Site and location

- country
- adm1
- adm2
- latitude
- longitude
- altitude_m

Crop and management context

- crop
- variety
- planting_date
- harvest_date
- production_system (e.g., rainfed, irrigated)
- irrigation (yes/no or categorical)
- fertilization (type or yes/no, when available)

Experimental and observational context

- observation_date
- season

- year

6. Conclusion

Through these coordinated actions, CIP has transitioned from ad hoc legacy data management to a structured, FAIR-by-design approach across crop breeding, agronomy, plant health, and biodiversity domains. By cleaning and FAIRifying existing datasets in CIP's Dataverse, adopting CGIAR Core Sets of Variables, and embedding these standards into future data collection workflows, CIP is ensuring that its datasets are interoperable, reusable, and AI-ready. This work directly supports CGIAR system-wide data governance objectives, strengthens cross-domain integration, and positions CIP's data assets as long-term resources for scientific analysis, decision support, and digital innovation.