Cross-Group Data Readiness Checklist (International Collaboration)

Purpose: A single, comprehensive checklist for each group in a (collaborative or not, potentially internationa) project to ensure data, documentation, protection, quality, and publication readiness.

😀 Document everything! Everything!!! 😎

A. Project & Dataset Metadata

☐ **Project identifiers present –** Project name, WP/task, contact/owner, contributors with ORCID.

- e.g., OrganoRice • WP2 • Contact: Dr. Xxx (ORCID: 0000-0000-1234-5678), contact data

☐ **Dataset title & abstract –** Clear, English title + 3–6 sentence abstract describing scope, time, location.

- e.g., "Field-level rice yields in Can Tho (2024), plot polygons and harvest records"

☐ **Keywords & geography –** Discipline keywords, study area, CRS if spatial.

- e.g., Keywords: rice, yield, organic, Mekong Delta • CRS: EPSG:4326 (European Petroleum Survey Group)

B. Documentation Essentials

☐ **README completed –** Explains dataset scope, folder structure (raw/interim/processed/docs/code), formats, CRS, license.

- e.g., README.md with sections 1–8 per template; UTF-8 CSV; ISO 8601 dates

☐ **Data dictionary / codebook –** Variables, units, allowed values, missing codes, collection instrument, QC rules.

- e.g., data_dictionary.csv & codebook.md covering FIELD_ID, YIELD_KG_M2, FARM_SIZE_HA, etc.

☐ **Versioning & changelog –** File naming with date/version; CHANGELOG records edits; Git repo if possible.

- e.g., yield_2024-08-31_v1.2.csv • CHANGELOG.md with entries for fixes

C. Structure, Formats & Conventions

☐ **Folder structure applied –** raw/ (immutable), interim/, processed/, docs/, code/.

- e.g., raw/surveys_2024.csv; processed/yield_per_m2_2024_v1.0.csv

☐ **Tabular format & encoding –** CSV, comma delimiter, dot decimal, UTF-8, header row in English.

- e.g., "FIELD_ID,YIELD_KG_M2,HARVEST_DATE"; 0.42; dates as 2024-08-31

☐ **Spatial data conventions –** CRS declared; area computations in projected CRS; units documented.

- e.g., plots.gpkg with layer metadata; area_m2 field derived in UTM 48N

D. Data Protection & Legal

☐ **Sensitive data classified –** Fields marked as non-sensitive / sensitive / special category; actions documented.

- e.g., PII removed from public tables; keyfile stored separately and encrypted

☐ **Consent & lawful basis –** Consent forms and/or legal basis recorded; DUA/DPIA if applicable.

- e.g., Signed consent forms stored in restricted drive; DUA template completed

☐ **Anonymization/pseudonymization –** Direct IDs removed; indirect identifiers generalized if re-identification risk.

- e.g., GPS generalized to 2-decimal degrees for public version

E. Storage, Backup & Security

☐ **3-2-1 backups documented –** 3 copies, 2 media, 1 offsite; restore tested; schedule noted.

- e.g., Primary: server; Secondary: external drive; Offsite: institutional cloud; last restore test: 2025-09-01

☐ **Access control & encryption –** Role-based access, MFA; encryption at rest and in transit for sensitive data.

- e.g., Restricted folder with MFA; keyfile in encrypted vault

☐ **Device hygiene –** No sole reliance on USB sticks; offboarding/revocation process documented.

- e.g., Departed student access revoked on 2025-07-15

F. Data Quality & Processing

☐ **Provenance captured –** Who/when/where/how; instrument models; SOP versions.

- e.g., Harvest weight via calibrated scale XYZ; SOP v2.1; collector: Le Thi A.

☐ **Raw vs. clean separation –** Raw are immutable; cleaning scripts versioned; outputs labeled by pipeline version.

- e.g., scripts/clean_yield_v0.3.R produces processed/yield_v0.3.csv

☐ **Validation & QC checks –** Range, domain, referential integrity, geospatial sanity (CRS/datum).

- e.g., Flag YIELD_KG_M2 > 3; verify FIELD_ID exists in plots table

G. Spatial / GPS Specifics (if applicable)

☐ **CRS & area correctness –** CRS explicitly stated; area computed in projected CRS; units consistent.

- e.g., Area_m2 computed in UTM zone 48N; yield = weight_kg / area_m2
- Initially define Reference System in the GIS you use (QGIS, ArcGIS, ...)

☐ **Location privacy –** Assess re-identification risk from precise coordinates; generalize if needed.

- e.g., Publish centroids rounded to 3 decimals instead of exact farm boundaries
- e.g., anonymize data so that individual farmers can not be derived from them

☐ **Sensor/imagery linkage –** Document satellite/source (e.g., Sentinel-2), acquisition dates, indices (NDVI/NDRE).

- e.g., Sentinel-2 tiles 2024-08-15; NDVI computed; cloud cover < 20%
- e.g., do the satellite data fit to what you need (TOA, BOA, Surface Reflectance)? Document any pre-processing (how do you get around cloud cover?)

H. Publication & Sharing

☐ **Repository chosen –** Generalist (Zenodo/Dryad/Figshare) or discipline-specific (e.g., GBIF).

- e.g., Zenodo community: OrganoRice; embargo until article acceptance

☐ **DOI / PID strategy –** DOI reserved/assigned; ORCIDs included for authors; dataset citation drafted.

- e.g., Zenodo DOI reserved; authors listed with ORCID; citation string in README
- Example
    1. Harald Puhl: https://orcid.org/0000-0002-6759-8700

2. Lutz Weihermüller: https://orcid.org/0000-0003-1991-7735

☐ **Data Availability Statement (DAS) –** Drafted per target journal (PLOS/Nature); references repository record.

- e.g., Nature-style DAS referencing Zenodo DOI and access conditions
- Example InnovAfrica: https://datadryad.org/dataset/doi:10.5061/dryad.dv41ns213

I. Interoperability & Internationalization

☐ **Language & encoding –** English for headers/metadata; UTF-8; glossary for local terms.

- e.g., Variable names in English; Vietnamese terms explained in README

☐ **Locale pitfalls avoided –** Decimal dot; ISO dates; no thousand separators; documented units (SI).

- e.g., 0.56 not 0,56; 2024-09-01; mass in kg; area in m^2

☐ **Licensing declared –** Clear license (CC BY 4.0 / CC0) and any restrictions stated.

- e.g., CC BY 4.0 badge in README and repository form

J. Reproducibility Package

☐ **Executable scripts/notebooks –** R/Python scripts or notebooks with instructions; minimal environment info.

- e.g., R script + sessionInfo(); Python requirements.txt; run.sh

☐ **Minimal test data –** Small sample to verify pipeline without sensitive content.

- e.g., A 10-row anonymized sample CSV in /docs/examples

☐ **Outputs & logs –** Key derived tables/figures reproducibly generated; logs/errors captured.

- e.g., processed/yield_summary_v1.0.csv; logs/clean_yield_2024-09-01.txt

K. Governance, Roles & Risk

☐ **Roles assigned –** PI, data steward, WP leads, publication lead; review/approval flow.

- e.g., Data steward approves README/data dictionary before upload

☐ **Risk register –** Top risks (loss, legal, skills) and mitigations listed.

- e.g., Risk: CRS mistakes → Mitigation: peer QC checklist before release

☐ **Final release review –** Peer review using this checklist; sign-off recorded.

- e.g., Signature/date lines; reviewer name and ORCID