Sensitive Data Classification – Template

Mark each dataset and field:
◻ Non-sensitive (open)
◻ Sensitive (restricted)
◻ Special category (strictly controlled)

- Direct identifiers: names, phone, email, address, national ID → remove or store separately.
- Indirect identifiers: GPS + small sample, rare categories → consider aggregation/generalization.
- Special categories: health, biometrics, religion, political views, genetics → avoid or strictly control.

Actions depending on type:
- Pseudonymize
- Encrypt
- Separate keyfiles: Using separate keyfiles for data protection enhances security by adding another required component for access, making brute-force attacks significantly more difficult and allowing for multi-factor authentication methods.
- DUA (Data Use Agreement: a formal, written, contractual agreement that establishes specific ways data may be used and how it must be protected.)
- DPIA (Data Protection Impact Assessment: a process to identify and mitigate privacy risks before they occur when processing personal data. It is required for processing activities likely to result in a high risk to individuals and involves describing the processing, assessing necessity and proportionality, identifying risks, and planning safeguards.. Mandatory for processing that is likely to result in a high risk to individuals);
- Retention & deletion schedule.

Project README file – Template

````
# README – Dataset Template (OrganoRice)

**Title:**
**Version:**
**Date:**
**Contact/Owner:**
**Contributors (ORCID):**

## 1. Description
- Project/work package context, scope, hypotheses
- Methods overview (SOP versions), instruments, calibration
- Study area, timeline

## 2. Files & Structure
```
/raw        # immutable raw inputs
/interim    # temporary/transformed
/processed  # analysis-ready outputs
/docs       # README, codebook, DMP, licenses
/code       # scripts, notebooks
```
- File list with short descriptions

## 3. Data Dictionary / Codebook
- See `data_dictionary.csv` and/or `codebook.md`

## 4. Formats & Conventions
- Encoding: UTF-8
- CSV with comma as delimiter, dot as decimal; ISO 8601 dates (YYYY-MM-DD)
- CRS for spatial data (e.g., EPSG:4326 or projected CRS)

## 5. Quality & Provenance
- Data lineage (sources, transforms), QC rules, known limitations
- Outlier handling policy, missing value codes

## 6. Access & Licensing
- Access level (open/restricted/controlled)
- License (e.g., CC BY 4.0, CC0)

## 7. Citation
- DOI for dataset (if assigned)
- How to cite

## 8. Acknowledgements
- Funding, partners, facilities
````

Storage & Backup Checklist – Template

Device, Storage & Backup Checklist

- MFA (Multi Factor Authentication) on all accounts; full-disk encryption enabled.
- 3-2-1 backups: 3 copies, 2 media, 1 offsite.
- Document backup frequency and test restores quarterly.
- No sole reliance on USB sticks; avoid single-laptop risk.
- Maintain access logs; revoke promptly when staff leave.

Data Availability Statement (DAS) – Templates

PLOS-style (example)

All relevant data are within the manuscript and its Supporting Information files. Where ethical or legal restrictions apply (e.g., personal data), de-identified data and metadata are available from the corresponding author upon reasonable request under a data use agreement.

Springer Nature / Nature-style (example)

The datasets generated during and/or analysed during the current study are available in the Zenodo repository, https://doi.org/XX.XXXX/zenodo.XXXXXXX, under the CC BY 4.0 license. Where data include sensitive information, access is restricted and can be granted by the Data Access Committee upon request; metadata record is openly available.

Custom (restricted data with public metadata)

Metadata describing the study design, variables, and methods are openly available at DOI:XX.XXXX. Due to privacy considerations and legal obligations (GDPR/PDPD), the underlying data cannot be made publicly available; controlled access can be requested via a data use agreement.

Data Management Plan (DMP) – Template

Project: OrganoRice (or related agriculture project)
Version: v1.0 | Date:
Contact/Owner:

1. Project Overview & Data Summary

• Objectives, work packages (WPs), data producers, stakeholders.
• Data types: tabular (CSV), spatial (GeoPackage/GeoTIFF), imagery (Sentinel-2), survey data, lab measurements.
• Expected volume, formats, and update frequency.

2. FAIR Principles

2.1 Findable

• Metadata standards (Dublin Core, DDI, Darwin Core, ISO 19115).
• Persistent identifiers (DOIs via DataCite/Zenodo; ORCID for authors).
• Naming conventions and versioning.

2.2 Accessible

• Access conditions (open/restricted/embargo/controlled).
• Repositories (Zenodo, Dryad, Figshare; institutional; GBIF for biodiversity).
• Authentication/authorization where applicable.

2.3 Interoperable

• Open, machine-readable formats (UTF-8 CSV, GeoPackage, GeoTIFF).
• Controlled vocabularies and codes; CRS definitions.
• APIs/OGC services if relevant.

2.4 Reusable

• Licensing (CC BY 4.0 / CC0 where appropriate).
• Rich metadata, methodology, README, codebooks/data dictionaries.
• Provenance and quality statements.

3. Roles & Responsibilities

• Data steward, PI, WP leads, IT support; review cycles and approvals.

4. Ethical, Legal & Sensitive Data

• GDPR & Vietnam PDPD alignment.
• Consent, anonymization/pseudonymization, DPIA where needed.
• Data use agreements (DUA).

5. Storage, Backup & Security

• 3-2-1 backups; encryption at rest/in transit; role-based access; MFA.
• Device hygiene, logging/auditing, incident response.

6. Processing & Reproducibility

• Raw/interim/processed foldering; version control (Git).
• Pipelines (R/Python/Jupyter), QC/QA, change logs.

7. Archiving & Preservation

• Long-term formats, repository preservation policies; retention and deletion schedule.

8. Budget & Resources

• Storage, compute, personnel (data steward), training; repository fees if any.

9. Risk Register

• Risks (data loss, legal non-compliance, skill gaps) and mitigation.

Data Dictionary (Data Points) – Template

table_name, field_name, label, description,
data_type, unit, allowed_values, missing_value_code, format,
example, collection_method, instrument_model, qc_rules, foreign_key, notes

plots,FIELD_ID,Field identifier,Unique identifier for field/plot,
string,,,NA,,
VN-CT-001,Survey,GPS receiver,Must be unique; not null,,

yield,YIELD_KG_M2,Yield (kg/m^2),Harvested yield normalized by area,
float,kg/m^2,,,0.00,
0.45,Weighing + area survey,Scale + tape,Range 0–5; flag values > 3 for
review,plots.FIELD_ID,Ensure area measured from surveyed polygon

table_name	field_name	label	description	data_type	unit	allowed_values	missing_value_code	format	example	collection_method	instrument_mod
plots	FIELD_ID	Field identifier	Unique identifier for field/plot	string			NA		VN-CT-001	Survey	GPS receiver
yield	YIELD_KG_M2	Yield (kg/m^2)	Harvested yield normalized by area	float	kg/m^2			0.00	0.45	Weighing + area survey	Scale + tape

Codebook (Survey) – Template

variable | question_text | type | values/codes | missing_code | notes

RESP_ID | Respondent identifier | string | — | NA | Stored separately from PII mapping |

AGE | What is your age? | integer | 18–99 | -99 | Validate range |

FARM_SIZE_HA | Farm size (ha) | float | — | NA | Check units and decimals |

PRACTICE_ORGANIC | Organic practices used | categorical | 0=no, 1=yes | 9 | Align with OrganoRice definitions |