With funding from the:
Federal Ministry of Research, Technology and Space

KIAG Transparency first

OrganoRice

Research Data for International Collaboration

„Data Accessibility and Archiving“

A lecture from the OrganoRice project

organorice.org

KIAG: Harald Puhl, Tom Anders

h.puhl@ki-ag.com / t.anders@ki-ag.com

CTU, Can Tho
08. November 2025

With funding from the:
Federal Ministry of Research, Technology and Space

Introduction

Data Accessibility & Archiving

Why is it important?

For students:
Increases chances of publishing internationally and building a professional reputation.

For staff:
Helps meet funder/journal requirements and protects your institution's research credibility.

For everyone:
Prevents data loss, saves time, supports collaboration with international partners and aligns with legal requirements.

With funding from the:
Federal Ministry of Research, Technology and Space

Introduction

Why is it important now?

- Vietnam's research output is growing rapidly.
- Vietnam currently implements new data protection rules (PDPL as of 1. January 2026).
- Funding agencies (EU, the World Bank, and regional ASEAN initiatives) increasingly require proper data management plans (DMPs) and data sharing according to the FAIR principles.
- Projects cannot achieve their full potential when data is not stored safely, not shared openly, or not understandable to international partners.
- Journals often reject manuscripts unless a data availability statement (DAS) and proper archiving are included.

With funding from the:
Federal Ministry of Research, Technology and Space

Learning Objectives

- Recognize data as an enabling infrastructure for international collaboration.
- Understand existing legal regulations and requirements.
- Understand challenges in international collaboration: language, formats, infrastructure.
- Understand international expectations like FAIR, journal policies, and funder requirements.
- Identify common pitfalls and how to avoid them.
- See the benefits: how good data practices improve your career prospects, strengthen collaboration, and safeguard your work.

With funding from the:
Federal Ministry of Research, Technology and Space

Motivation

Utilize Fair principles in your research:

F - Findable

A - Accessible

I - Interoperable

R - Reusable

With funding from the:
Federal Ministry of Research, Technology and Space

Topics

- Master Data & Documentation
- Data Protection, Security & legal aspects
- Data quality & processing
- Data publication & sharing
- Wrap-up

With funding from the:
Federal Ministry of Research, Technology and Space

Master Data & Documentation (MDD)

Learning objectives

Design master data and identifiers to link activities across work packages (WP) and institutions.

Apply documentation practices (data dictionary, README, codebook) and metadata standards (measurement units).

With funding from the:
Federal Ministry of Research, Technology and Space

MDD: Core Content

- Identifiers & naming: persistent IDs for plots, farms, fields, sensors; file naming with dates (YYYY-MM-DD), version (v1.1), and scope.
- Data dictionary: variable name, label, type, allowed values, units (SI), missing-value codes, collection instrument, QC rules.
- README: project scope, collection methods, instruments, units, coordinate system, contact, license.
- Metadata: choose discipline-appropriate standards (Dublin Core for general; DDI for surveys; Darwin Core for biodiversity; ISO 19115/19157 for geospatial).
- Repositories: structure folders (raw/, interim/, processed/, docs/), maintain a CHANGELOG; use Git for version control.

With funding from the:
Federal Ministry of Research, Technology and Space

MDD: README file 1/2

README – Dataset Template (OrganoRice)

Title:

Version:

Date:

Contact/Owner:

Contributors (ORCID):

1. Description

- Project/work package context, scope, hypotheses

- Methods overview (Standard Operating Procedures), 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

With funding from the:
Federal Ministry of Research, Technology and Space

MDD: README file 2/2

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

With funding from the:
Federal Ministry of Research, Technology and Space

MDD: Data Dictionary

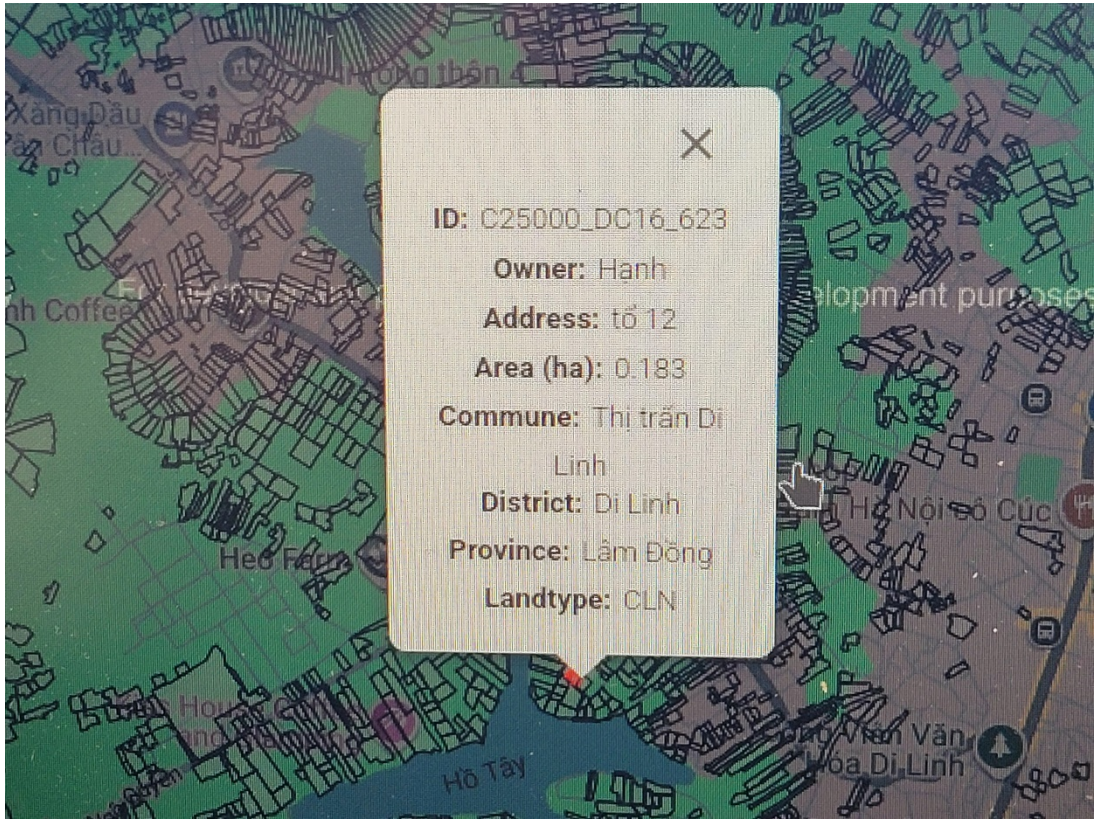
table_name, field_name, label, description,
data_type, unit, allowed_values, missing_value_code, format,
example, collection_method, instrument_model, qc_rules, foreign_key, notes

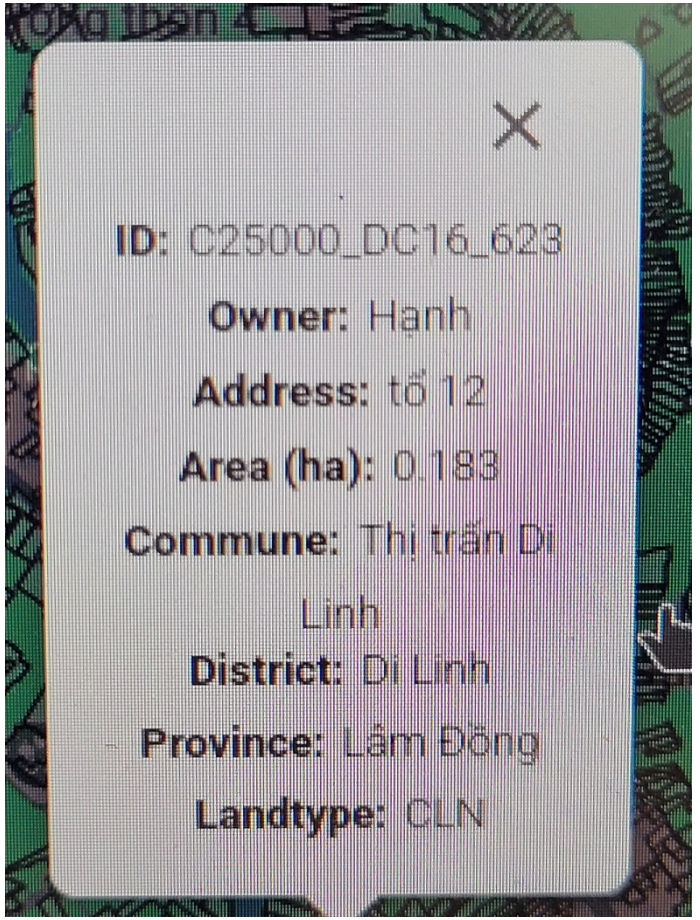
plots,FIELD_ID,Field identifier,Unique identifier for field/plot,
string,,,NA,,
VN-CT-001,Survey,GPS receiver,Must be unique; not null,,

yield,YIELD_KG_M2,Yield (kg/m²),Harvested yield normalized by area,
float,kg/m²,,,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

With funding from the:
Federal Ministry of Research, Technology and Space

MDD: Việt Nam Example – FIELD_ID





With funding from the:
Federal Ministry of Research, Technology and Space

MDD: Codebook - Survey Variables

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 |

With funding from the:
Federal Ministry of Research, Technology and Space

MDD: Pitfalls

Lost context: Dataset with 'Q12' and 'YLD' but no codebook; three years later, variables are uninterpretable.

Unit confusion: Yield recorded as tons/ha vs. kg/m² – mixing leads to KPI errors for 'yield per square meter'.

With funding from the:
Federal Ministry of Research, Technology and Space

Data Protection, Security & Safety (DSS)

Learning objectives

- **Differentiate critical/sensitive data from non-sensitive data**
- **Apply GDPR and Vietnam PDPL legal basics**
- **Implement proportionate safeguards (role-based access, encryption, backups).**

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: Why data protection?

Data protection is regarded as a fundamental right, based in the

- right of privacy and the
- right of reputation, dignity and honour.

- Safeguarding of the reputation of (Vietnamese and EU) institutions
- The EU and Vietnam have introduced dedicated and similar data protection laws (GDPR and PDPL) to implement these rights.

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: Legal groundwork

EU - General Data Protection Regulation (GDPR)

Regulation (EU) 2016/679

https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

Vietnam – Personal Data Protection Law (PDPL)

Law No. 91/2025/QH15

https://thuvienphapluat.vn/van-ban/EN/Bo-may-hanh-chinh/Law-91-2025-QH15-Personal-Data-Protection/665440/tieng-anh.aspx

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: Legal groundwork

EU Regulation GDPR (EU) 2016/679 & **Vietnam** PDPL (Law No. 91/2025/QH15)

- Always ask participants for explicit consent before collecting or using personal data. Consent must be informed, voluntary, and recorded.
- Collect only the information truly necessary for your research. Avoid collecting identifiable data unless essential.
- Remove or replace personal identifiers whenever possible. Use coded IDs instead of names.
- Store data securely (encrypted or password-protected). Only authorized team members should have access.
- Clearly tell participants what data you collect, why, and how long you'll keep it. They can request access or deletion of their data at any time.

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: What is data protection? Personal data!

These laws provide extensive rights for individuals regarding their personal data. The key factors are:

- Personal data can only be used if a legal basis exists (e. g. consent)
- Only the data necessary for the specific purpose may be collected
- Persons have the right to be informed which data is collected about them
- Persons can ask for data to be corrected or deleted
- Sensitive categories of data (e. g. health data) cannot be stored unless for a limited set of exceptions

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: But what is personal data?

Personal Data

direct identifiers (names, phone, email, address)

indirect identifiers (GPS + small sample)

special category (health, biometrics, religion, political views, genetic)

Special attention must be paid to vulnerable groups.

- Children
- Elderly people
- Sick or disabled people
- Minorities
- People in a dependent relationships (e. g. farmers)

They are more prone to misuse of their personal data or less able to know and enforce their rights.

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: Data security and safety

Data protection can only be achieved if data is

- (security) trusted, available and secure
- (safety) used in the right way

How they interact:

- You can be secure but unsafe (e.g., perfectly locked data used to target vulnerable users).
- You can be safe-minded but insecure (great policies, but an open S3 bucket).

Secure the data and safeguard its use across the full project lifecycle!

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: What is data security and safety?

Data security =

keeping data confidential, intact, and available to the right people only.

Think locks, alarms, and guards

Data safety =

making sure data is handled and used in ways that don't cause harm to people, systems, or society.

Think rules of use, ethics, and safeguards against misuse

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: In more detail - Data Security

Confidentiality:

Only authorized parties are able to view or use the data (access rights, anonymization, encryption)

Integrity:

Data is not modified, corrupted or deleted

Accessibility:

Data is stored safely and is available even in case of failure of single devices

What you should avoid:

- **Do not** share accounts
- **Do not** save access/login-details or even data on public devices
- **Do not** rely on a single data storage when collecting, maintaining, and analyse data

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: In more detail - Data Safety

Prevent harm:

Policies, processes, and technical measures in place during data collection, processing, and usage

Responsible use:

Focus on minimizing and de-identifying sensitive data. This honors consent and adds guardrails

Rules & Behavior:

In practice it combines the rules (clear roles, audits, impact assessments) with your behavior (anonymization, data minimization, bias testing) during entire data lifecycle

What you should avoid:

- **Do not** collect data without clarifying data protection measures
- **Do not** collect data without consent
- **Do not** collect data from sensitive categories or vulnerable groups
- **Do not** collect unnecessary data
- **Do not** keep data forever

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: Data security and safety in a table

Aspect	Data Security	Data Safety
Core goal	Prevent **unauthorized access**/alteration/outage (CIA triad)	Prevent **harmful outcomes** from authorized collection, processing, and use
Typical threats	Hackers, malware, insider theft, ransomware	Privacy violations, misuse of sensitive data, harmful model outputs, bias, unsafe automation
Common controls	Encryption, access control, network segmentation, MFA, patching, monitoring/IR, backups	Data minimization, anonymization, purpose limitation, consent, policy & governance, bias testing, red-teaming, human-in-the-loop, DLP, audit trails
Compliance focus	Security standards (e.g., ISO 27001, SOC 2)	Privacy/ethics and sector rules (e.g., GDPR principles, AI safety policies)
Example	Encrypt a database and restrict admin access	Remove PII from training data; block a model from generating harmful actions using that data

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: Anonymization & Pseudonymization of data?

Pseudonymization means an anonymous alias is used instead of the actual personal data, but the alias is linked to the full data.

Example:

Peter Smith is assigned the alias "person123". Data in public sources only show person123, only a person with access to the full data can see the link to Peter Smith.

While pseudonymized data can be used without restriction, the link between the pseudonyms and the personal data cannot and is protected by data protection laws.

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: Anonymization & Pseudonymization of data?

Anonymization means the data is changed so much that identifying the person is no more possible even with additional information.

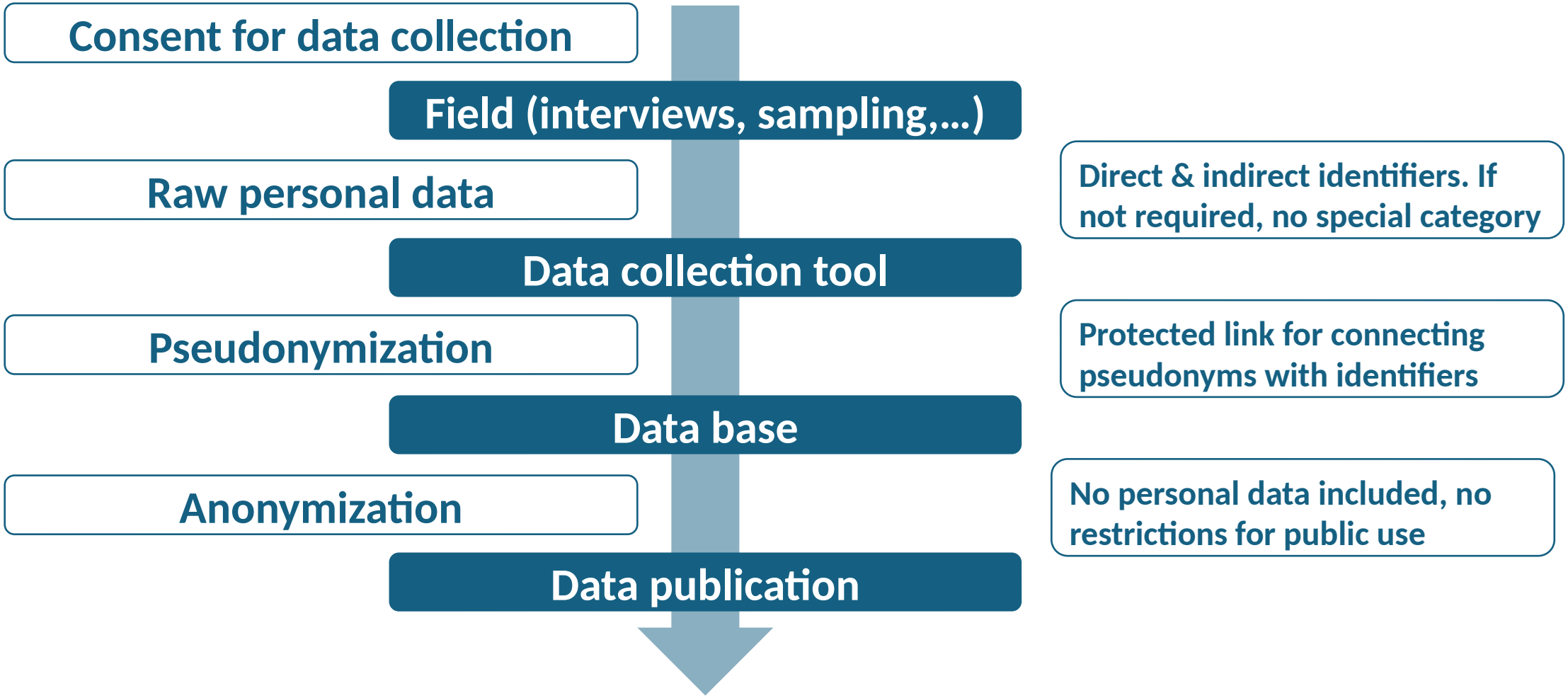
Example:

A person entered the birthdate April 19, 1984. This is changed to "born in the 1980s", making it impossible to identify the person.

Anonymized data is no longer regarded "personal information" and can be used without being restricted by data protection laws.

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: Data protection process



With funding from the:
Federal Ministry of Research, Technology and Space

DSS: Clear so far. What else?

<u>It is crucial to classify the data that will be collected</u>

- **Non-sensitive data** (may be open).

  Example: Country, Region, City
- **Sensitive data** (restricted/controlled).

  Example: Name, email
- **Special category** (strictly controlled or not publishable).

  Example: Health, Biometrics, Religion, Political views

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: Consent for data use and storage

Before data is being collected, **participants should be informed about the data use and storage.**

Documentation creates the legal basis for using this data.

<u>What needs to be documented:</u>

- Contact details of the person/organisation controlling the data
- Description of the data that is collected
- Purpose of collecting, using, and storing the data
- Who will use this data (local and international)
- How long the data will be stored
- Information about the participants rights
- Contact to request Information and access

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: How to proceed in practise?

For example: you want to collect data about the number of children in households:

- only collect name, address, email and phone number of a parent (for potential questions), but not the bank account of the parents or the phone numbers or messenger contacts of children, their age, names or birthdates, etc.).
- generate a random unique ID for each entry which does not contain personal data so you can anonymize the data

Non-sensitive data: Region, City, maybe District, number of children

Sensitive data: street/number, email, phone number

Special category: not used

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: How to achieve data protection in practice?

To protect the data, a reliable storage infrastructure must be provided, and best practices be used.

- access concept (define user rules and access rights, strong passwords, MFA)
- use encryption for storage and transmission, store keyfiles separately
- redundant hardware with monitoring of data integrity (e. g. hash checks, audit trails)
- Working backup setup, 3-2-1 rule (3 copies, 2 media, 1 offsite)

Not to forget factors like device hygiene - only use the software needed for the project, no private data and 3rd party apps on the devices

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: Achieve data protection - Challenges

Data Safety Data security measures

From a data safety perspective, access has to be strictly limited while data security can profit from extensive monitoring of data/data movements.

The motto is therefore **"as open as possible, as closed as necessary".**

MP must specify access controls for reading/writing data as w tion and deletion.

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: Takeaways and what happens, if…

Takeaways:

- Non-compliance with data protection laws has legal consequences in both countries.
- Publishing person-identifiable data without consent or proper legal basis is a policy and compliance failure and can lead to repercussions by authorities.

<u>For your project:</u>

- take the time to plan your projects thoroughly and
- think about the legal requirements at every step along the way.

With funding from the:
Federal Ministry of Research, Technology and Space

DSS: Pitfalls & Stories

Otherwise, small omissions and errors can lead to delays and/or disciplinary measures. Examples:

- A table (CSV file) containing survey entries including the names and telephone numbers of all participants is accidentally uploaded to a web site and is thus available publicly worldwide. A partner in the EU blocks the publication of the survey and demands a corrective action plan. Possibly a complaint is filed with the data protection supervisor, and a fine could be issued.
- A map of harvest yields is created, including high-precision GPS data of single farms - making these farms (and farmers) identifiable. This is a good example about how seemingly non-sensitive data can lead to data protection issues through indirect identifiers. To remedy this problem, the GPS data could be generalized (e. g. showing a larger area instead of an exact pinpoint), or the access to the map is strictly restricted to a defined set of users with the consent of the involved farmers.

With funding from the:
Federal Ministry of Research, Technology and Space

Relevance for Agriculture

Dataset	Personal/sensitive?	Comments and practical handling
Farmer lists (names, contacts)	Personal data	Collect on a clear purpose; get consent/notice; limit fields; set retention tied to project
GPS coordinates (plot-level)	Sensitive personal data	When linkable to a person/household (location = sensitive) → stronger duties. Use coarser grids for analysis exports
Agricultural production data	Personal data	If linked to a farmer/plot; aggregate or pseudonymize for sharing/publication
Satellite imaging for land use	Usually non-personal	becomes personal/sensitive if tied to a named owner or precise home/plot location. Mask/blur identifying features before sharing. (Inference from definitions + sensitive location rule.)
Terrain inspection (photos, videos, drone)	Personal data	Can capture sensitive data (faces, homes, exact location). Blur faces/houses; store raw only where necessary; restrict access
Field soil samples	Non-personal	unless linked to an identifiable farmer/plot; keep linkage in a separate, access-controlled table

With funding from the:
Federal Ministry of Research, Technology and Space

Data Quality & Processing (DQP)

Learning objectives

Understand key terms: Data provenance; ETL/ELT; QC/QA; Outlier; Schema validation; Reproducibility; Jupyter; R scripts; Git; Issue tracker.

Design reproducible pipelines; preserve raw data; document transformations.

Define quality dimensions (accuracy, completeness, consistency, timeliness, validity).

With funding from the:
Federal Ministry of Research, Technology and Space

DQP: Quality dimensions

Accuracy: The degree to which data correctly reflects the real-world objects or events it represents.
Example: When a fertilization trial in Vinh Long is logged as “compost only,” the record should exactly match the protocol actually used in the field, in line with OrganoRice’s traceability-focused data capture.

Completeness: The extent to which all required data elements and records are present.
Example: For every experimental field on the project’s data map (green = plant protection, red = fertilization, purple = rice quality), include all required attributes (e.g., location, dates, inputs, irrigation, yield, lab results) for each site—none missing.

Consistency: The alignment of data values across datasets and systems without contradictions.
Example: Plot IDs and province names must be identical across the opportunity map, meeting minutes, and reports—for example, Vinh Long, An Giang, and Dong Thap should appear the same everywhere.

Timeliness: How current the data is and whether it’s available when needed.
Example: Public pages should be updated promptly as events unfold—e.g., reflecting Tra Vinh’s joining on January 15, 2025 and keeping the ISOP 2025 page current ahead of the Nov 3–6, 2025 symposium at Can Tho University.

Validity: The extent to which data conforms to expected formats, rules, and constraints.
Example: Measurements and metadata should conform to agreed formats/units and pass schema checks—for instance, parameters in the irrigation strategy report must use valid units and ranges appropriate for the Mekong Delta.

With funding from the:
Federal Ministry of Research, Technology and Space

DQP: Core Content

- Context capture: who/when/where/how; instrument models & calibration; environmental conditions; SOP (standard operating procedures) versions.
- Raw vs. clean: Raw are immutable; cleaning scripts are versioned; outputs are labeled by pipeline version.
- Validation: range checks, domain checks, referential integrity (foreign keys), geospatial sanity like CRS (coordinates reference system), datum.
- Tools: OpenRefine for profiling & transforms; R/Python scripts; notebooks for literate workflows (Jupyter, deepNote); GitHub/GitLab for version control; QGIS/GDAL for spatial QC.

With funding from the:
Federal Ministry of Research, Technology and Space

DQP: Context Capture

Essential to capture the **context** of data capture to make capture **reproducible** and **comparable**

- who/when/where/how
- instrument models & calibration (e.g. temperatures)
- environmental conditions (leave cover impacts GPS)
- SOP (standard operating procedure) versions which were followed

This information is also required for data publication

With funding from the:
Federal Ministry of Research, Technology and Space

DQP: Raw vs. Clean

Only if **raw** data are **preserved** they can be **checked** for capturing or processing pipeline errors **later**.

- **Raw** data are immutable – **never changed**
- Cleaning **scripts** and the whole pipeline are **versioned** (if there are no scripts, the cleaning descriptions are versioned)
- Cleaned **outputs** are labeled by **pipeline version**

With funding from the:
Federal Ministry of Research, Technology and Space

DQP: Validation

There are different types of checks depending on the type and complexity of data.

- Range checks (numeric: water > 100°)
- Domain checks (non-numeric: type of fertilizer element of list?)
- Referential integrity (foreign keys, associated to the right farm)
- Geospatial sanity like CRS (coordinates reference system)
- Date (does the date correspond)

With funding from the:
Federal Ministry of Research, Technology and Space

DQP: Ex. Range and Domain Checks

- **Province code**: must be one of {An Giang, Đồng Tháp, Vĩnh Long, …}; anything else is flagged.
- **Treatment type**: value ∈ {Control, Compost, Compost+Biochar, Organic NPK, …} from the project codebook.
- **Soil pH**: must be within a realistic agronomic range for Mekong Delta paddies (e.g., 3.5–9.0).
- **Yield (t/ha)**: between 0 and an upper bound set by agronomists (e.g., ≤15); negatives or extreme outliers are rejected.
- **GPS coordinates**: must lie inside Vietnam/Mekong Delta bounding boxes; points outside are out-of-domain.

With funding from the:
Federal Ministry of Research, Technology and Space

DQP: How to implement checks?

- Use **declarative constraints** (e.g., SQL CHECK (ph BETWEEN 3.5 AND 9.0); CHECK (moisture <= 14)).
- Make **ranges conditional** on context (province/season/variety): CASE WHEN province='An Giang' THEN rate BETWEEN 0 AND 180 ELSE rate BETWEEN 0 AND 200 END.
- **Normalize units** before checking; reject or convert mixed units.
- NOT IN/lookup joins against **reference tables** (province IN ref_province)
- **ETL rules** that fail or quarantine rows when values fall outside the domain
- Decide **failure policy**: hard fail (reject row) vs quarantine + review for near-misses/outliers.
- **Log violations** with field, value, expected range, and key (plot_id, sample_id) to speed triage

Databases and programming help a lot!

With funding from the:
Federal Ministry of Research, Technology and Space

DQP: Tools

If you do not feel like database programmings, there are tools available to help you:

- Profiling & transforms
- R/Python scripts
- Notebooks for literate workflows
- Version control
- Spatial QC

With funding from the:
Federal Ministry of Research, Technology and Space

DQP: Brief Toolset Description

- **OpenRefine (profiling & transforms)**: Point-and-click profiling with facets and clustering reveals typos, near-duplicates, and category drift; transformation recipes (GREL) let you standardize columns at scale.
  Best for: fast, one-off or semi-structured cleaning of CSV/Excel with an auditable undo/redo history.

- **R/Python scripts**: Programmatic checks (e.g., pandera, pydantic, Great Expectations) enforce schemas, ranges, and cross-field rules, and can automate fixes.
  Best for: repeatable, scalable QC pipelines that run on servers/CI and integrate with ETL and databases.

- **Notebooks (Jupyter, Deepnote)**: "Literate" workflows mix narrative, code, and visuals to explore anomalies, show why rules exist, and keep QC logic transparent.
  Best for: exploratory profiling, prototyping validation rules, and communicating QC findings to collaborators.

- **GitHub/GitLab (version control)**: Stores code, schemas, and QC configs with full history, reviews, and issues; CI/CD can run tests on every change to catch regressions.
  Best for: collaboration, traceability, and automating QC checks on commit/merge.

- **QGIS/GDAL (spatial QC)**: Visual and batch tools for CRS fixes, topology rules, and spatial tests (in-bounds checks, duplicates by distance, slivers, self-intersections).
  Best for: validating and correcting geospatial data, especially coordinates, geometries, and spatial relationships.

With funding from the:
Federal Ministry of Research, Technology and Space

DQP: Surface Data are Important

Special case: GPS/surface data

- CRS awareness (WGS84 EPSG:4326 vs. projected CRS); area computation in appropriate projection; differential GPS error; canopy/multipath issues.
- Central KPIs depend on surfaces, proper surface calculation is hence extremely important

Example:
Yield per m^2: ensure plot area measured consistently (surveyed polygon vs. assumed rectangular field).

With funding from the:
Federal Ministry of Research, Technology and Space

DQP: Pitfalls & Stories

- Over-cleaning: Deleting 'outliers' that were valid hill-slope plots with lower yields; KPI bias introduced
- CRS mismatch: Mixing lat/long with UTM meters – area and distance wrong by orders of magnitude

With funding from the:
Federal Ministry of Research, Technology and Space

Data Publication & Sharing (DPS)

Learning Objectives

- Identify and select suitable repositories (generalist, discipline-specific, or institutional).
- Draft a Data Availability Statement (DAS) that complies with funder and journal requirements.
- Understand and apply licenses (CC0, CC BY) to clarify permissions for reuse.
- Assign Persistent Identifiers (PIDs) such as DOIs to datasets and connect them to researcher identities (ORCID).

With funding from the:
Federal Ministry of Research, Technology and Space

DPS: Choosing a Repository

Selection criteria

Disciplinary relevance: Does the repository fit your field?
Don't upload biodiversity data to a physics repository - it won't reach your target users.

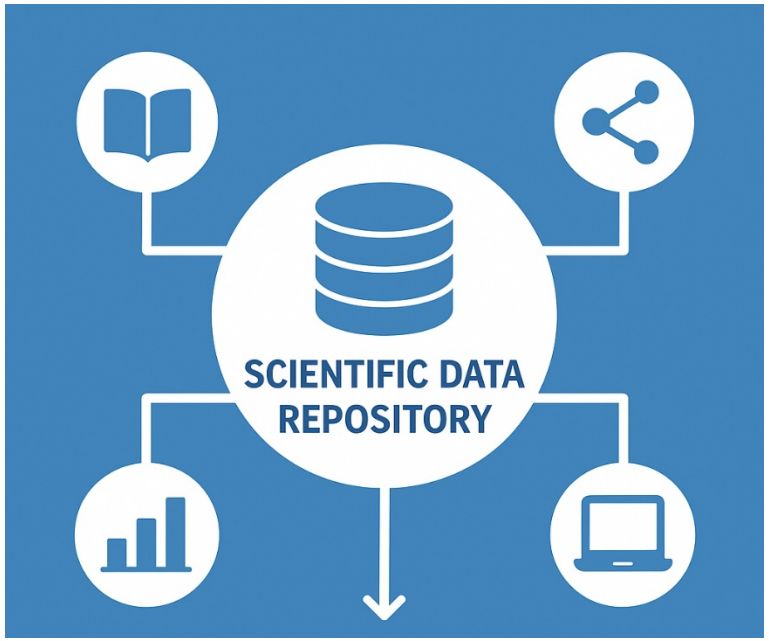
Trustworthiness: Recognized by funders/journals, long-term preservation guaranteed.
Avoid saving data only on Google Drive or USBs.

Visibility: Indexed in Google Dataset Search and OpenAIRE.
Don't assume "any website" makes data visible.

Metadata quality: Requires structured metadata for discoverability.
Don't skip metadata fields, it makes your data invisible.

Cost & access: There are free repositories and fee-based ones.
Check before upload

A **repository** is a virtual archive to **collect, disseminate and preserve** scientific output like articles and datasets. You can make them available to other researchers based on your personal reference. (OpenAIRE, 2025)



Repository: Articles & datasets for shared & open online access

With funding from the:
Federal Ministry of Research, Technology and Space

DPS: Choosing a Repository

Generalist:

Zenodo (EU-funded, free, integrates with GitHub), Dryad, Figshare.

Discipline-specific:

GBIF (biodiversity), ICPSR (social sciences), GenBank (genomics).

Institutional:

Some Vietnamese universities are developing repositories. If not available, international repositories are recommended.

With funding from the:
Federal Ministry of Research, Technology and Space

DPS: Persistent Identifiers & Visibility

Persistent identifiers (PIDs) are unique references that help identify researchers, research institutions and digital objects. (Wiley, 2024)

Why is it persistent?

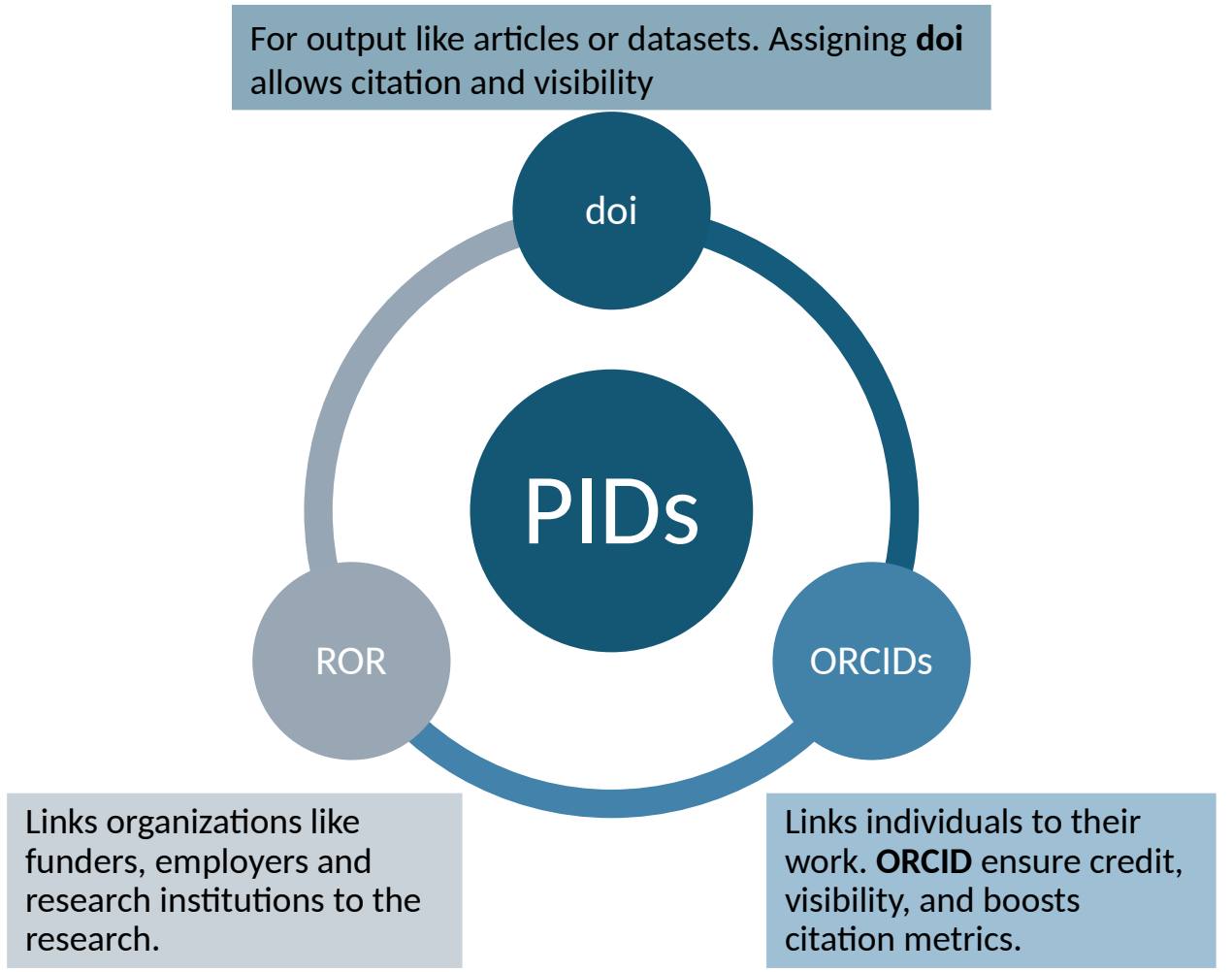
- It identifies the object even if it is moved or changed.

Example:

A Vietnamese PhD student linked her Zenodo dataset DOI to her ORCID.

Her work can be easily located and cannot be confused with others that share similar names.

Within a year, her dataset was reused and cited by a European project - boosting her profile and credibility.



With funding from the:
Federal Ministry of Research, Technology and Space

DPS: Data Availability Statement

What is a DAS?

Data **A**vailability **S**tatement

- Short comment if the research data (evidence) is made available and where it can be accessed.

Why is it required?

Journals like *Nature*, *PLOS*, *Springer Nature* require clear data availability.

With funding from the:
Federal Ministry of Research, Technology and Space

DPS: Data Availability Statement

Examples:

Poor DAS: "Data available upon request." (often rejected by journals).

Good DAS: "All **raw survey data and R scripts** are available in **Zenodo** (**DOI: 10.5281/zenodo.1234567**) under **CC BY 4.0** license."

Includes: the data type, the repository incl. IDs and the license for further use.

With funding from the:
Federal Ministry of Research, Technology and Space

DPS: Data Formats & Interoperability

Essential for easy access and international cooperation

File formats:

- Use open, machine-readable formats (CSV, TXT, NetCDF). Avoid proprietary-only formats (e.g., Excel with macros)

Encoding:

- Always use UTF-8 for text

Numbers & dates:

- Use dot decimal (1.234), comma-separated CSV, and ISO 8601 dates (YYYY-MM-DD)

Documentation includes:

- *README file*
- *Codebook/metadata* (definitions of variables, units, methods)
- *CRS (Coordinate Reference System)* information for spatial data

With funding from the:
Federal Ministry of Research, Technology and Space

DPS: Licensing Data

Different License types

- **CC0:** Complete waiver - anyone can reuse data without restrictions
  → Best for maximum openness
- **CC BY:** Requires attribution - the name of the creator and attribution parties, a notice and a link to the material must be indicated
  → Best to protect researcher recognition while enabling reuse
- Other licenses exist, providing stricter regulations

You want your research to be reused?

Choose CC0 or CCBY & consider the repository's requirements

Concerns

- Researcher often worry about plagiarism or misuse

Solution

- Licensing clarifies rights and provides international legal backing

You are still not sure which License to use?

Try this tool:

SCAN ME

creative commons

With funding from the:
Federal Ministry of Research, Technology and Space

DPS: Pitfalls & Stories

- **Locale trap:** A Vietnamese team saves CSV with *semicolon delimiters and comma decimals* → U.S. collaborator's pipeline fails. Weeks lost.
- **Missing context:** Raster dataset uploaded without CRS (coordinate system). U.S. partners cannot georeference. Data becomes useless.
- **Positive story:** A Can Tho research group shared rice cultivation data on Zenodo with DOI + CC BY license. Result: cited in an EU Horizon proposal, leading to follow-up collaboration.

With funding from the:
Federal Ministry of Research, Technology and Space

DPS: DIY – Prepare your publication

Hands on exercise (10min)

You are a researcher from Vietnam. You want to publish data from your soil samples on the topic of organic agriculture practices.

Example doi: 10.3390/rs16173221, orcid: 0009-0003-8752-8106

Provide the following information:

- Choose a Repository
- Choose a license
- Write a DAS
- (Generate a dataset citation)

Bonus:

- Create a ORCID for yourself https://orcid.org/

With funding from the:
Federal Ministry of Research, Technology and Space

Wrap-up

- Create Master data for your datasets (MDD)
- Provide a data dictionary, README & codebook and comply to metadata standards (MDD)
- Non-compliance with data protection laws has legal consequences in both countries. (DSS)
- Publishing person-identifiable data without consent or proper legal basis is a policy and compliance failure and can lead to repercussions by authorities. (DSS)
- Define Quality dimensions and document data processing (DQP)
- Always provide context for all data (DQP)
- Use repositories and make your data visible and identifiable (DPS)
- Provide clear permissions for reuse (licencing) and comply to publisher requirements (DPS)

With funding from the:
Federal Ministry of Research, Technology and Space

Questions?

OrganoRice Project

organorice.org

Harald Puhl, Tom Anders

h.puhl@ki-ag.com / t.anders@ki-ag.com