Session 2

Data Preparation & Metadata Editor Basics

Session 2: Data Preparation & Metadata Editor Basics

LNADA Capacity Building Training

Lao Statistics Bureau

February 23, 2026

Why Data Preparation Matters

The Metadata Editor is NOT a data editing tool

- Data must be clean BEFORE documentation
- Use spreadsheet software or statistical packages for data cleaning
- Metadata Editor: documents and publishes what you have
- Data quality issues → metadata quality issues

Step 1: Organise Your Files

Standardised Directory Structure

```
STUDY_ID_2026/
├── README.txt (description)
├── data/
│   ├── hld_2026_v01.dta
│   ├── ind_2026_v01.dta
│   └── (all data files in one folder)
├── documentation/
│   ├── questionnaire.pdf
│   ├── codebook.xlsx
│   └── technical_notes.docx
├── external_resources/
│   ├── field_report.pdf
│   ├── thumbnail.jpg
│   └── analytical_report.docx
└── archive/
    └── (raw, unprocessed data — never used for documentation)
```

Step 2: Preserve Original Files

Always Keep Unaltered Copies

- **Golden rule**: Never overwrite original data files
- Create working copies if you need to edit
    - e.g., `data_original.dta` + `data_working.dta`
- The Metadata Editor is NOT a backup system
- You upload files to the Editor — they are NOT backed up automatically
- Archive approach: compress old versions, store securely

Quality Checks: Identifiers & Duplicates

Data Integrity Foundations

Unique Identifiers:
- Every observation must have a unique ID (household ID, person ID)
- No missing values in the ID field
- No duplicate ID values
- No special characters in IDs (use only: letters, numbers, underscore)

Duplicate Observations:
- Check: are there duplicate rows?
- Example: "Person 001 appears twice in data" = problem
- Use statistical package to identify: `duplicates report id` (Stata) or `duplicated()` (R)

File Relationships:
- Multiple files? Ensure they link properly on common ID
- Example: household file + individual file should both have `hld_id`
- Test: can you merge without duplicating records?

Quality Checks: Variable Names & Labels

Variable Names:
- Must be unique (no duplicates)
- Use standard naming: `income_annual`, `age_years`, `region_code`
- Avoid spaces, special characters (except underscore)
- Maximum 32 characters (for SPSS/Stata compatibility)

Variable Labels:
- Every variable should have a human-readable label
- Example: `q01_occupation` → "What is your primary occupation?"
- Labels appear in metadata and catalogs — make them clear

Quality Checks: Value Labels & Ranges

Value Labels (Categorical Data):
- Define all possible values
- Example: `region_code` = {1: "Vientiane Capital", 2: "Champasak", ...}
- Document missing codes: {99: "Not stated", 98: "Not applicable"}

Value Ranges (Numerical Data):
- Check min/max values are reasonable
- Example: age should be 0–120, not 0–999
- Detect outliers: are there extreme values that look like data entry errors?

Quality Checks: Data Types & Missing Values

Data Types:
- Declare each variable as: numeric, string, date, boolean
- Example: Income must be numeric, not stored as text "5000.00"
- Date variables: use ISO 8601 format (YYYY-MM-DD)

Missing Values:
- Document how missing data is coded
- Common missing codes: {99, 999, -999, blank, NA}
- Flag as missing in metadata, not as valid value
- Calculate: % missing per variable — high % suggests data quality issue

Quality Checks: Sample Weights & Privacy

Special Variables and Sensitive Data

Sample Weights:
- If survey: what is the sampling design?
- Probability proportional to size (PPS)?
- Stratified sampling? Cluster sampling?
- Document the weight variable and the design
- "All estimates using this data should employ the sample weight"

Privacy & Confidentiality:
- **CRITICAL**: Remove direct identifiers before sharing
  - Names, addresses, phone numbers, email addresses
  - National ID numbers, passport numbers
  - Biometric data (fingerprints, photos)
- Keep anonymised codes (hld*id*, *person*id, etc.)
- Aggregate sensitive variables (age → age groups, income → income brackets)
- Check: is this data safe to share with researchers?

Quality Checks: Data Compression

File Formats and Optimization

Recommended Formats:
- **Stata** (.dta) — best for statistical data with labels and value labels
- **SPSS** (.sav) — alternative, widely compatible
- **CSV** (.csv) — universal, but loses labels and data types (need external codebook)
- **Excel** (.xlsx) — only for small files or documents (not ideal for data)

Compression:
- Large files? Use .zip or .7z compression
- Example: 100 MB .dta → 20 MB when compressed
- Metadata Editor supports: .zip, .rar, .7z, .gz
- Users download compressed, extract on their end

"Bad Data" vs. "Good Data" Example

Real-World Comparison

Aspect	Bad Data	Good Data
ID variable	Missing values, duplicates	Unique, no gaps, no nulls
Variable names	`X1`, `Q_001_1A_new`, special chars	`age_years`, `income_annual`
Variable labels	None, or cryptic "Q1"	"What is your primary occupation?"
Value labels	Codes only (1, 2, 3)	1: "Yes", 2: "No", 99: "Missing"
Missing values	Mixed (blank, -999, NA, "N/A")	Consistent: 99 = missing, documented
Data types	Age stored as text: "25.0"	Age as numeric
Identifiers	Name: "John Smith" still in data	Removed; only ID: "001" remains
Duplicates	Person 001 appears 3 times	Each ID appears exactly once

Data Preparation Checklist (1/2)

Before You Upload to Metadata Editor

- [ ] **Files organised** in standardised folder structure
- [ ] **Originals preserved** in archive or backup location
- [ ] **Unique identifiers** checked: no missing, no duplicates
- [ ] **No duplicate observations** (checked with statistical software)
- [ ] **Variable names** are unique, clear, follow naming convention
- [ ] **Variable labels** written for every variable
- [ ] **Value labels** assigned (for categorical variables)

Data Preparation Checklist (2/2)

- [ ] **Missing values** documented and consistently coded (e.g., 99)
- [ ] **Value ranges** verified (no unreasonable min/max)
- [ ] **Data types** correct (numeric, string, date, etc.)
- [ ] **Sample weights** assigned and documented (if sample data)
- [ ] **Direct identifiers removed** (names, addresses, contact info)
- [ ] **Files compressed** if large (> 50 MB)

Supported Data Types — Quick Recap

8 Data Types Supported by Metadata Editor

Type	Standard	Example	Today's Focus
Documents	Dublin Core / MARC21 / BibTex	Reports, publications, PDFs	✓ S3 Exercise 3a
Microdata	DDI Codebook 2.5	Household survey, census	✓ S3 Exercise 3b
Indicators	WB schema	GDP, poverty rate	✓ S4 Exercise 4a
Geographic	ISO 19139	Province boundaries, maps	S4 if time
Statistical Tables	SDMX	Aggregate summary tables	S4 if time
Images	IPTC	Photographs, infographics	S4 if time
Videos	Schema.org	Training videos, YouTube	S4 if time
Scripts	Dedicated schema	R/Python analysis scripts	S4 if time

Metadata Editor Login

Training Instance: editor.lsb.lao-stat.de

Training Accounts:

Role	Email	Password
Admin	trainingadmin@lsb.gov.la	lsbLaoStatII@0223
Regular User	traininguser@lsb.gov.la	lsbLaoStatII@0223

First-time Login:
1. Open browser → editor.lsb.lao-stat.de
2. Click "Sign In"
3. Enter email and password
4. You will see the **Dashboard** with project list

User Management: Create a New Account

Site Administration → Users → Add User

Step-by-step for Exercise 1:

1. Log in as **trainingadmin@lsb.gov.la**
2. Click **Site Administration** (top menu)
3. Select **Users** from left sidebar
4. Click **Add User** button (top right)
5. Fill form:
   - **Full Name:** Your name
   - **Email:** Your email address (will be username)
   - **Password:** Create a strong password (8+ chars)
   - **Role:** Select "Administrator" (for this training)
6. Click **Create**
7. **Verify:** Log out, then log in with your new account

Exercise 1: Create Your User Account

Hands-On (10 minutes)

1. Open browser → editor.lsb.lao-stat.de

2. Sign in as **trainingadmin@lsb.gov.la** (password: lsbLaoStatII@0223)

3. Go to **Site Administration → Users → Add User**

4. Fill in: your name, email, password, role = "Administrator"

5. Click **Create**

Exercise 1: Verify Your Account

1. Log out (click your name in top right → Logout)

2. Log in with your **new** email and password

3. Verify you see the Dashboard

If stuck: Raise your hand. Trainer will help.

NADA Training Instance

Create Your NADA Account

NADA = the catalog where published data appears

Training Instance: nada.lsb.lao-stat.de

Default Training Accounts (same as Metadata Editor):

Role	Email	Password
Admin	trainingadmin@lsb.gov.la	lsbLaoStatII@0223
Regular User	traininguser@lsb.gov.la	lsbLaoStatII@0223

Exercise 1b: Create NADA Admin Account

Hands-On (5 minutes)

1. Open browser → nada.lsb.lao-stat.de
2. Sign in as **trainingadmin@lsb.gov.la** (password: lsbLaoStatII@0223)
3. Go to **Site Administration → Users → Add User**
4. Fill in: your name, email, password, role = **Administrator**
5. Click **Create**
6. Log out → Log in with your **new** NADA account
7. Verify you see the NADA Dashboard

Exercise 1c: Create Your API Key

Required for publishing from Metadata Editor to NADA

Why? The Metadata Editor needs an API key to authenticate with NADA when publishing.

1. Log in to NADA with your **new** account
2. Click your name (top right) → **Profile**
3. Or go directly to: nada.lsb.lao-stat.de/index.php/auth/profile
4. Find the **API Keys** section
5. Click **Generate New API Key**
6. **Copy and save** the key somewhere safe (you will need it in Session 4)