Pilot Project Report

João P. Donadio

2025-08-09

NDIS Database Analysis Overview

Completed NDIS notebook review and improvements as a pilot project for Lasisi Labs.

Total time spent: ~10 hours

Start date: Aug. 07, 2025

End date: Aug. 09, 2025

Work Log

1. Initial Setup (1 hour)

- Kick-off meeting
- Forked repository and created `pilot_Joao` branch
- Set up clean VSCode environment
- Tested end-to-end notebook execution

2. Code Review & Improvements (3 hours)

Issues identified:

- Path handling not OS-agnostic
- Noisy logging and code chunks results

Solutions implemented:

- Refactored paths using `pathlib`
- Added progress logging and error retry wrappers
- Adjusted codes for reproducibility

3. Data Wrangling (3 hour)

- Transformed raw scraped data → tidy format
- Validated consistency across timepoints
- Output clean `ndis_fixed.csv`

4. Visualization (1 hour)

- Created plots for data visualization
- Added annotations and styling

5. Documentation (2 hours)

- Restructured Quarto document flow
- Wrote this report and README files

Pipeline

The project analyzes the growth of the FBI's National DNA Index System (NDIS) by extracting and processing historical statistics from 300+ Wayback Machine snapshots (2010-2025). The pipeline systematically collects, cleans, and visualizes data on offender, arrestee, and forensic DNA profiles across U.S. jurisdictions.

Key Pipeline Components

1. **Data Collection**
    - Wayback Machine API queries to identify relevant snapshots
    - Parallel downloading of HTML files with error handling
    - Validation of file integrity and completeness
2. **Data Extraction**
    - Era-specific parsing (pre/post-2010 formats)
    - Metadata extraction including report dates
    - Jurisdiction name standardization
3. **Data Processing**
    - Type conversion and quality flagging
    - Temporal feature engineering (year/quarter/era)
    - Derived metrics (total profiles, forensic ratios)
4. **Analysis & Visualization**
    - Time-series growth trends by profile type
    - Geospatial mapping of participation
    - Interactive data exploration tools

Technical Highlights

- **Robust Architecture**: Parallel processing with automatic retries
- **Temporal Handling**: Specialized parsers for different report eras
- **Validation**: Cross-language checks (Python/R) and statistical validation
- **Reproducibility**: Version-controlled outputs with metadata

Outputs

- Cleaned longitudinal dataset (2010-2025) -> data/v1.0/ndis_data_v1.0.csv
- Summary statistics tables
- Interactive and Static visualizations

SDIS Pipeline Proposal

Objective: Process a messy master sheet of 50 states' DNA data to derive a clean, well-documented dataset with a standardized `n_total_estimated` variable, accounting for arrestees, offenders, and forensic profiles.

Estimated time to complete: 7 hours.

1. Data Architecture & Repo Tidy-Up (20 min)

Actions:

- Identify the "canonical" location for the raw master sheet (e.g., `data/raw/sdis_master.csv`).
- Document file paths and naming conventions in the Quarto notebook.

2. Input Data QC + Diagnostics (2h40)

Actions:

- **Initial Audit**:
  - Flag missing/inconsistent values (e.g., states with only `n_arrestees` but no `n_total`).
  - Document assumptions (e.g., "If only `n_offenders` and `n_arrestees` exist, sum them for `n_total_estimated`").
- **QC Logic Implementation**:
  - Create rules for `n_total_estimated` based on available columns (see decision tree below).
  - Add a column `data_source` to track how each state's total was derived (e.g., "sum of arrestees/offenders", "reported total minus forensics").
- **Diagnostic Visualization**:
  - Generate a **heatmap** showing data availability (arrestees/offenders/forensics/total) per state.
- **Decision Rules for `n_total_estimated`**:
  1. If `n_total` is reported **and** matches `n_arrestees + n_offenders` → Use `n_total`.
  2. If `n_total` is reported **but** includes forensics → Subtract forensics: `n_total_estimated = n_total - n_forensic`.
  3. If only `n_arrestees` and `n_offenders` exist → Sum them.
  4. If `n_total` conflicts with partial data (e.g., Connecticut's `n_total = n_offenders` but arrestees exist) → Use `n_total` and note discrepancy.
- **Deliverable**:
  - Cleaned dataset with `n_total_estimated` and QC heatmap.

3. State-by-State Totals Pipeline (3h)

Approach:

- Process states **programmatically** where possible (e.g., using R loops), but handle edge cases manually (e.g., Missouri, Connecticut).
- For each state:
  - Apply QC logic above.
  - Record exceptions in the notebook (e.g., "Assumed Missouri's `n_total` includes arrestees as the difference from offenders").
- **Documentation**:
  - Add a **narrative section** explaining edge cases and assumptions.
  - Use code comments and markdown to justify decisions (e.g., "Forensic profiles excluded per client request").
- Create a **summary visualization** (Map with bar chart) showing:
  - States with/without arrestees/offenders/forensics.
  - Final `n_total_estimated` per state.

Deliverable: Transparent, reproducible pipeline with descriptions for each state.

4. Document Structure and Data Exportion (1h)

Actions:

- Structure the Quarto notebook as:
  1. **Introduction**: Project goal and `n_total_estimated` definition.
  2. **Data Cleaning**: QC logic, edge cases, and heatmap.
  3. **Pipeline**: State-by-state processing steps.
  4. **Output**: Final dataset and visualization.
- Ensure reproducibility and a clear communication.
- Freeze and save files.

Deliverables:

- Professional, self-contained report (sdis_summary.qmd)
- Final data to `data/v1.0/sdis_data_v1.0.csv`.

FOIA Demographic Data Processing Plan

Objective: Harmonize 7 states' DNA demographic data (PDFs/CSVs) into a clean long-form dataset with provenance tracking and visualizations.

Estimated time to complete: 15 hours.

1. Metadata Schema Design (1h)

Actions:

- Create a **concise metadata table** (`foia_state_metadata.csv`) documenting for each state: state, report_levels, race_data_type, gender_data_type, total_profiles_source, notes.
- Design **column spec** for final dataset:

```
foia_combined <- tibble(
  state = character(),              # e.g., "California"
  offender_type = character(),      # e.g., "Juvenile", "Adult"
  variable_category = character(), # "gender", "race", or "total"
  variable_detailed = character(), # e.g., "Female", "Hispanic"
  value = numeric(),                # Numeric value (count/percentage)
  value_type = character(),         # "count" or "percentage"
  value_source = character(),      # "reported" or "calculated"
  year = integer()                  # e.g., 2023
)
```

Deliverable: Clear schema documentation in Quarto notebook.

2. State-by-State QC & Parsing (8h)

Processing Pipeline for Each State:

1. **Load & Standardize**:
   - Convert PDF tables → CSV (OCR already done)
   - Standardize terms (e.g., "Arrested offender" → "Arrestee")
2. **Handle Edge Cases**:
   - **California**: Add "Unknown" race bucket to reconcile totals
   - **Indiana**: Calculate counts from percentages (only % provided)
   - **Texas**: Infer male counts from female-only reporting
   - **Nevada**: Map "flags" → "profiles" terminology
3. **Reconcile Totals**:
   - Verify `sum(demographic_counts) == total_profiles`
   - Add "Unknown" categories where shortfalls exist
   - Handle rounding (e.g., "<1%" → 0.5%)
4. **Tag Data Provenance**:
   - Add `value_source` column ("reported" vs. "calculated")

Deliverable: 7 cleaned state datasets with QC checks documented in `output/foia/state_foia_data.csv`

3. Harmonization Script (3 hrs)

Actions:

- Merge all states into **long-form** `foia_combined.csv`
- Calculate **combined totals** where needed.
- Derive missing percentages/counts.

Deliverable: Single harmonized dataset with all transformations tracked in the qmd.

4. Validation Figures (2 hrs)

Visualizations:

1. **State-Level**:
    - Bar charts showing race/gender proportions per state
2. **Combined Heatmap**:
    - Grid showing which states report arrestees/offenders/combined
    - Color-coded by data completeness

Deliverable: 2-3 key plots embedded in Quarto doc.

5. Final Prose & Export (1h)

Actions:

- Write **methods summary** explaining all data imputations/transformations
- Freeze final dataset as `data/v1.0/foia_combined_v1.0.csv`
- Structure notebook

Deliverable: Publication-ready Quarto document.