# **README — OECD AI Enterprise Survey (3NF)**

## **Purpose**

This dataset contains enterprise-level OECD AI Survey data structured for relational database integration and analytical modeling.  
It is designed to serve as a **normalized data source** next to the n\_load\_db environment.

The processing script (data\_wrangeling\_oecd\_data.py) performs:

* Extraction of all OECD survey tables from the raw CSV.
* Cleaning of irregular delimiters and formatting inconsistencies.
* Removal of percent symbols (%) while preserving numeric and contextual information.
* Normalization into **3NF** across meta, dimension, and fact tables.
* Export of UTF-8 encoded CSVs ready for SQL database import.

## **Input**

**Source:** OECD\_surveys.csv  
Contains multiple survey tables on AI usage, governance, and investment indicators.  
Each table corresponds to one OECD question block (e.g., *“Table E1 Q1 – How important are AI applications to your enterprise's core business processes?”*).

### **Extraction Logic**

* Tables are detected by headers beginning with "Table " (case-insensitive).
* Each table extends until the next "Table " line or end of file.
* Irregular spacing and delimiters are normalized using regex-based parsing.
* Parentheses are retained (e.g., 12 (40.0)), while % symbols are removed.

## **Output Files (3NF)**

All files are UTF-8 encoded with headers and no index columns.  
Each source table generates **three CSVs**:

| **File** | **Table** | **Description** |
| --- | --- | --- |
| Table\_E1\_Q1\_meta.csv | OECD\_Meta | Survey-level metadata (table name, source file) |
| Table\_E1\_Q1\_dimensions.csv | OECD\_Dimensions | Contextual attributes (country, enterprise size, sector, etc.) |
| Table\_E1\_Q1\_facts.csv | OECD\_Facts | Observations or indicators linked to dimensions and meta table |

## **Database Schema (3NF)**

CREATE TABLE OECD\_Meta (

table\_id TEXT PRIMARY KEY,

table\_name TEXT,

source\_file TEXT

);

CREATE TABLE OECD\_Dimensions (

dimension\_id TEXT PRIMARY KEY,

Country TEXT,

"Enterprise size" TEXT,

Sector TEXT

-- Additional contextual attributes as detected

);

CREATE TABLE OECD\_Facts (

fact\_id SERIAL PRIMARY KEY,

table\_id TEXT REFERENCES OECD\_Meta(table\_id),

dimension\_id TEXT REFERENCES OECD\_Dimensions(dimension\_id),

indicator TEXT,

value TEXT

);

## **Relationships**

OECD\_Meta (1) ───< OECD\_Facts (many)

OECD\_Dimensions (1) ───< OECD\_Facts (many)

* **Each OECD table (meta)** may contain many fact rows.
* **Each dimension record** (e.g., CAN / 50–249 / ICT) may appear across multiple indicators.
* **Facts** hold the actual numeric or categorical values (e.g., “Training of employees – Yes”, “73.3”).

## **Loading Examples**

### **MySQL**

LOAD DATA INFILE '/path/Table\_E1\_Q1\_meta.csv'

INTO TABLE OECD\_Meta

FIELDS TERMINATED BY ',' OPTIONALLY ENCLOSED BY '"'

LINES TERMINATED BY '\n'

IGNORE 1 LINES;

Add referential constraints:

ALTER TABLE OECD\_Facts

ADD CONSTRAINT fk\_facts\_meta

FOREIGN KEY (table\_id) REFERENCES OECD\_Meta(table\_id);

ALTER TABLE OECD\_Facts

ADD CONSTRAINT fk\_facts\_dimensions

FOREIGN KEY (dimension\_id) REFERENCES OECD\_Dimensions(dimension\_id);

## **Validation Queries**

### **Referential Integrity**

SELECT COUNT(\*) FROM OECD\_Facts f

LEFT JOIN OECD\_Meta m ON f.table\_id = m.table\_id

WHERE m.table\_id IS NULL;

SELECT COUNT(\*) FROM OECD\_Facts f

LEFT JOIN OECD\_Dimensions d ON f.dimension\_id = d.dimension\_id

WHERE d.dimension\_id IS NULL;

### **Data Quality Checks**

| **Check** | **Result** |
| --- | --- |
| All tables have unique table\_id and dimension\_id | Done |
| Percent symbols removed, parentheses preserved | Done |
| No empty rows or columns in CSV exports | Done |
| UTF-8 encoding confirmed | Done |

## **Developer Notes**

* **Primary Keys:** table\_id (Meta), dimension\_id (Dimensions), fact\_id (Facts).
* **Foreign Keys:** OECD\_Facts.table\_id → OECD\_Meta.table\_id,  
   OECD\_Facts.dimension\_id → OECD\_Dimensions.dimension\_id.
* **Data Integrity:** maintained through controlled parsing, null filtering, and ID linkage.
* **Granularity:** Each fact row corresponds to a single data point from the OECD tables.
* **Extensibility:** New OECD survey rounds can be appended using the same schema.
* **Compatibility:** Fully aligned for relational import in PostgreSQL, MySQL, or SQLite.
* **Encoding:** All exports are UTF-8 with standard commas as delimiters.