# **README — OECD AI Survey Questionnaire Dataset**

## **Purpose**

This dataset represents the **OECD AI Survey Questionnaire structure**, designed for integration as a **metadata source** within the EUROSTAT\_load\_db relational database.

It contains parsed and normalized data describing the survey’s **sections**, **questions**, and **response options**, all stored in **3NF** for referential integrity and efficient querying.

The processing pipeline performs:

* Extraction and cleaning of raw OECD survey text (from Tabula or OCR).
* Detection of sections, questions, and bullet-style answer options.
* Classification of response types (Yes/No, Don’t Know, Multiple Choice).
* Normalization into 3NF: Section, Question, and Option tables.
* Export of UTF-8 encoded CSVs ready for database import.

## **Input**

**Sources:**

* OECD data.csv — Text-based OECD survey questionnaire (raw text format).
* tabula-OECD data.csv — OCR/Tabula multi-column extraction of the same questionnaire.

### **Parsing Logic**

1. **Section Detection** — Lines ending with “questions” (e.g., *“Screening questions”*).
2. **Question Detection** — Lines beginning with “Question” or standalone items not starting with “•”.
3. **Option Detection** — Lines starting with “•” interpreted as answer options.
4. **Response Type Detection** — Automatically inferred from text patterns (e.g., *Yes / No / Don’t know*).
5. **Cleaning** — Removal of commas, artifacts, and OCR noise while preserving parentheses and option groupings.

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

| **File** | **Table** | **Description** |
| --- | --- | --- |
| OECD\_3NF\_section.csv | Section | High-level grouping of survey questions |
| OECD\_3NF\_question.csv | Question | Individual survey questions linked to sections |
| OECD\_3NF\_option.csv | Option | Answer options linked to questions |

All files are **UTF-8 encoded**, contain headers, and exclude index columns.

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

CREATE TABLE Section (

SectionID TEXT PRIMARY KEY,

SectionName TEXT

);

CREATE TABLE Question (

QuestionID TEXT PRIMARY KEY,

SectionID TEXT REFERENCES Section(SectionID),

QuestionText TEXT,

ResponseType TEXT

);

CREATE TABLE Option (

OptionID TEXT PRIMARY KEY,

QuestionID TEXT REFERENCES Question(QuestionID),

OptionText TEXT

);

## **Relationships**

Section (1) ───< Question (many)  
 Question (1) ───< Option (many)

Each section (e.g., *Screening questions*) contains multiple questions.  
Each question may have multiple predefined response options.

## **Loading Examples**

### **MySQL**

LOAD DATA INFILE '/path/OECD\_3NF\_section.csv'

INTO TABLE Section

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

LINES TERMINATED BY '\n'

IGNORE 1 LINES;

Add referential constraints:

ALTER TABLE Question

ADD CONSTRAINT fk\_question\_section

FOREIGN KEY (SectionID) REFERENCES Section(SectionID);

ALTER TABLE Option

ADD CONSTRAINT fk\_option\_question

FOREIGN KEY (QuestionID) REFERENCES Question(QuestionID);

## **Validation Queries**

### **Referential Integrity**

SELECT COUNT(\*) FROM Question q

LEFT JOIN Section s ON q.SectionID = s.SectionID

WHERE s.SectionID IS NULL;

SELECT COUNT(\*) FROM Option o

LEFT JOIN Question q ON o.QuestionID = q.QuestionID

WHERE q.QuestionID IS NULL;

### **Primary Key & Encoding Checks**

| **Check** | **Result** |
| --- | --- |
| SectionID unique | ✅ |
| QuestionID unique | ✅ |
| OptionID unique | ✅ |
| UTF-8 encoding verified | ✅ |
| Referential integrity valid | ✅ |

## **Developer Notes**

**Primary Keys:** SectionID, QuestionID, OptionID  
 **Foreign Keys:** Maintain referential integrity between all layers.  
 **Granularity:** One row per option per question.  
 **Deterministic IDs:** OptionID generated from hash of (QuestionID + OptionText).  
 **Encoding:** All exports are UTF-8.  
 **Extensibility:** Supports future OECD survey versions and translations.  
 **Indexing:**

CREATE INDEX idx\_question\_section ON Question(SectionID);

CREATE INDEX idx\_option\_question ON Option(QuestionID);