## **Goal: Combine Annex C + Annex E inside your database**

have two types of normalized data:

| **Dataset** | **Origin** | **Content** | **Normalization Level** |
| --- | --- | --- | --- |
| **Questionnaire (Annex C)** | Parsed from OECD survey text | All survey questions, sections, and options | 3NF → Section, Question, Option |
| **Aggregated Results (Annex E)** | Extracted from OECD result tables | Aggregated percentages by country, sector, size, question | 3NF → TableMeta, Dimension, Fact |

join them through the **QuestionID** ↔ **TableID** mapping.

## **Step 1: Core Tables in Your Database**

should have (or create) these six tables:

-- From Annex C (Questionnaire)

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

);

-- From Annex E (Aggregated Results)

CREATE TABLE TableMeta (

TableID TEXT PRIMARY KEY, -- e.g. 'E.1', 'E.2'

QuestionID TEXT REFERENCES Question(QuestionID),

Title TEXT, -- 'How important are AI applications...'

Source TEXT -- e.g. 'OECD Annex E'

);

CREATE TABLE Dimension (

DimensionID TEXT PRIMARY KEY,

Country TEXT,

EnterpriseSize TEXT,

Sector TEXT

);

CREATE TABLE Fact (

TableID TEXT REFERENCES TableMeta(TableID),

DimensionID TEXT REFERENCES Dimension(DimensionID),

Indicator TEXT, -- e.g. 'Yes', 'No', 'Very important'

Value NUMERIC

);

## **Step 2: Establish the Mapping (Crosswalk)**

Create a **Crosswalk table** linking each question (from Annex C) to its result table(s) (Annex E).

CREATE TABLE QuestionTableMap (

QuestionID TEXT REFERENCES Question(QuestionID),

TableID TEXT REFERENCES TableMeta(TableID),

PRIMARY KEY (QuestionID, TableID)

);

## **Step 3: Relational Flow (3NF → Analytical Join)**

in SQL:

SELECT

s.SectionName,

q.QuestionText,

tm.TableID,

d.Country,

d.Sector,

d.EnterpriseSize,

f.Indicator,

f.Value

FROM Fact f

JOIN TableMeta tm ON f.TableID = tm.TableID

JOIN Question q ON tm.QuestionID = q.QuestionID

JOIN Section s ON q.SectionID = s.SectionID

JOIN Dimension d ON f.DimensionID = d.DimensionID

WHERE q.QuestionID = 'Q1'

ORDER BY d.Country, d.Sector, d.EnterpriseSize;

## **Step 4: Data Loading Flow**

| **Step** | **Script** | **Output** |
| --- | --- | --- |
| 1 | parse\_oecd\_csv\_3nf.py | section.csv, question.csv, option.csv |
| 2 | extract\_all\_oecd\_tables.py | E\_\*.csv tables (raw results) |
| 3 | normalize\_oecd\_tables\_3nf.py | TableMeta, Dimension, Fact CSVs |
| 4 | Load all CSVs into database |  |
| 5 | Populate QuestionTableMap crosswalk |  |

## **Step 5: Using the Combined Dataset**

Once loaded, you can:

* **Query metadata + results** (e.g., all OECD results for questions on “transparency”).
* **Compare** your internal survey’s answers to OECD aggregates by joining on QuestionID.
* **Generate dashboards** combining:  
  + OECD reference values (Annex E)
  + Your company/region responses (from your AI Compatibility Test)
  + Metadata (from Annex C)

Example query comparing internal vs OECD averages:

SELECT

q.QuestionText,

d.Country,

f.Value AS OECD\_Percent,

i.Value AS Internal\_Percent,

(i.Value - f.Value) AS Gap

FROM Fact f

JOIN QuestionTableMap m ON f.TableID = m.TableID

JOIN Question q ON m.QuestionID = q.QuestionID

JOIN Dimension d ON f.DimensionID = d.DimensionID

LEFT JOIN InternalResults i ON q.QuestionID = i.QuestionID AND d.Country = i.Country;