# **Data Project Plan: Benchmarking & Integration Database**

**Project Title:** Database: integration & benchmarking  
**Prepared by:** Vicky  
 **Date:** 10-09-2025

## **1. Project Goal**

To build a central relational database that consolidates benchmarking data across organizations, time periods, and regions and different data types, integrating information from both external data sources and client-submitted surveys (via Qualtrics).

This database will serve as the central place for all benchmarking activities, allowing consistent comparison, validation, and analysis across diverse datasets.

**2. Objectives**

1. **Integrate multiple data sources** — synthetic, research-based, industrial, and Kaggle datasets — into one standardized relational structure.
2. **Incorporate client submissions** from Qualtrics surveys for real-time analysis.
3. **Enable cross-sectional and longitudinal analysis** (across organizations, regions, and years).
4. **Standardize indicator definitions** and ensure consistent data quality across sources.
5. **Support analytics and reporting tools** (PUT IN FRONT-END, Power BI, Tableau, Python, R).
6. **Maintain data security, quality, and traceability** for all sources.
7. **3NF normalization is required!**

## **3. Data Sources Overview**

| **Source Type** | **Description** | **Example** | **Data Characteristics** | **Update Frequency** |
| --- | --- | --- | --- | --- |
| **Synthetic Data** | Artificially generated to simulate realistic benchmarking patterns | Generated in Python / R | Controlled quality, no confidentiality concerns | On demand |
| **Research Data** | From published academic or policy studies | OECD, World Bank, academic reports | High credibility, may require normalization | Annual / Ad hoc |
| **Industrial Datasets** | Proprietary or partner datasets from industry | HR analytics, ESG performance data | Real-world data, variable structure | Quarterly / Annual |
| **Kaggle Datasets** | Public datasets | Employee satisfaction, diversity, productivity | Publicly available, varying quality | Annual / Ad hoc |
| **Client Survey Data (Qualtrics)** | Client-submitted self-reported data | Qualtrics survey results | Real-time, self-entered | Continuous |

## **4. Proposed Database Structure**

| **Table Name** | **Description** | **Example Fields** |
| --- | --- | --- |
| UPON REQUEST DUE DO HAVING SEVERAL DATASETS | UPON REQUEST DUE DO HAVING SEVERAL DATASETS | UPON REQUEST DUE DO HAVING SEVERAL DATASETS |

**Relationships:**

UPON REQUEST DUE DO HAVING SEVERAL DATASETS

## **5. Geographic & Temporal Dimensions**

The database must support:

* **Regional granularity:** Country, region, subregion
* **Temporal granularity:** Year, quarter, or month
* **Cross-sectional views:** Comparing organizations or industries within a given year
* **Longitudinal views:** Tracking performance over time

## **6. Data Flow Overview**

1. **Wrangled data (Vicky):** Clean and standardize all external data sources → prepare benchmark tables.
2. **Qualtrics integration:** Import new client data submissions (API or CSV).
3. **Transformation & validation layer:** Align formats, apply data cleaning rules.
4. **Integration layer:** Merge and reconcile all data by indicator, time, and region.
5. **Analytical outputs:** Benchmark comparisons, scorecards, dashboards.

## **7. Qualtrics Survey Integration**

| **Aspect** | **Description** |
| --- | --- |
| **Input Format** | CSV export or API JSON |
| **Mapping** | Each survey question maps to an Indicator in the database |
| **Identifiers** | client\_id, survey\_id, survey\_year |
| **Validation** | Ensure correct data types, mandatory fields, and value ranges |
| **Automations** | Future integration: scheduled Qualtrics API pulls |

*Example Mapping:*

| **Qualtrics Field** | **Database Field** | **Description** | **Transformation** |
| --- | --- | --- | --- |
| Not yet determined | Not yet determined | Not yet determined | Not yet determined |
| Not yet determined | Not yet determined | Not yet determined | Not yet determined |
| Not yet determined | Not yet determined | Not yet determined | Not yet determined |

## **8. Data Wrangling Deliverables**

From the data wrangling (benchmarking) team → to the database developer:

| **Deliverable** | **Description** | **Format** |
| --- | --- | --- |
| **Data Dictionary** | Field names, definitions, data types, units, source | Excel |
| **Wrangled Data Files** | Cleaned benchmark datasets | CSV |
| **Field Mapping** | Qualtrics → Database mapping | Excel |
| **Transformation Documentation** | Steps used to clean and merge datasets | Word / Markdown |
| **Sample Queries** | Example analysis queries | SQL snippets |

### **10. Governance, Quality & Security**

| **Aspect** | **Description** |
| --- | --- |
| **Access Control** | Client-level survey data (from Qualtrics) is restricted to authorized users only. Each client’s submissions are isolated in the database, following role-based access policies. |
| **Data Provenance** | Every record (including Qualtrics submissions) is tagged with source\_id, load\_date, and submission\_id to track origin, version, and ingestion date. |
| **Quality Flags** | Automatic validation applied to all incoming data — including Qualtrics surveys — to detect missing fields, duplicates, or out-of-range responses. Invalid submissions are quarantined for review. |
| **Versioning** | Historical versions of survey submissions are maintained per load cycle. Updated or resubmitted surveys are version-controlled to preserve data lineage. |
| **Backup & Security** | Regular backups, encrypted storage, and secure authentication for both database and Qualtrics API integrations. |
| **Compliance** | All survey data handling follows GDPR and client data protection standards, with anonymization where applicable. |