Diseño de bases de datos y recolección de información

Seminario de Proyecto I (MCDI) — Unidad 4

Sergio Martín Nava Muñoz

2025-10-15

Propósito de la sesión

- Entender los **principios básicos para organizar datos** en un proyecto de investigación.
- Crear un **plan para recolectar información** de manera ordenada y ética.
- Preparar los elementos necesarios para la **Evidencia 4A** (base de datos) y presentación final.

Tip

¿Qué vamos a crear hoy?
1) Esquema de organización de datos (¿qué tablas necesito?).
2) Diccionario de variables (descripción de cada campo).
3) Plan de recolección de datos.
4) Una muestra pequeña de nuestra base de datos.

Pasos para organizar nuestros datos

1. **Identificar variables** → ¿Qué necesito medir para mi investigación?
2. **Diseñar estructura** → ¿Cómo organizo la información en tablas?
3. **Definir relaciones** → ¿Cómo se conectan las tablas entre sí?
4. **Establecer reglas** → ¿Qué datos son válidos y cuáles no?
5. **Documentar todo** → Crear un diccionario que explique cada variable.
6. **Planear recolección** → ¿De dónde y cómo obtendré los datos?
7. **Verificar calidad** → ¿Cómo me aseguro de que los datos sean confiables?

De variables a organización de datos

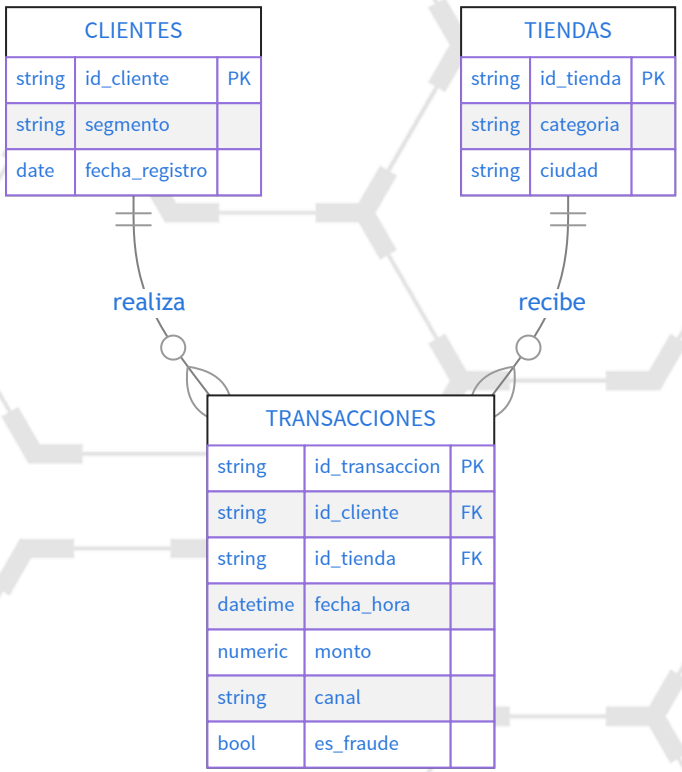
- **Variable dependiente**: lo que quiero explicar o predecir.
- **Variables independientes**: los factores que pueden influir.
- **Variables derivadas**: nuevas variables que calculo a partir de las anteriores.

Ejemplo (estudio de fraude en transacciones):
- Variable dependiente: `es_fraude` (sí/no).
- Variables independientes: `monto`, `canal`, `tienda`, `cliente`, `fecha_hora`.
- Variables derivadas: `hora_del_dia`, `dia_semana`, `monto_promedio_ultimo_mes`.

Grupos de información
- `clientes` (información personal)
- `tiendas` (información comercial)
- `transacciones` (cada compra)

¿Cómo se relacionan?
- Un cliente puede hacer muchas transacciones
- Una tienda puede recibir muchas transacciones
- Cada transacción pertenece a un cliente y una tienda

Esquema simple de organización



> Este es un diagrama básico. Adáptalo a tu proyecto específico.

Conceptos clave para organizar datos

- **Identificadores únicos**: cada fila en una tabla debe tener un código único.
- **Referencias entre tablas**: usar códigos para conectar información relacionada.
- **Tipos de datos apropiados**: números para cantidades, texto para nombres, fechas para tiempo.
- **Reglas de validación**: establecer qué valores son aceptables.

Reglas básicas

- Cada tabla debe tener un identificador único principal.
- Usar códigos simples para conectar tablas (no texto largo).
- Mantener los identificadores estables (no cambiarlos después).

Diccionario de datos - ¿qué es cada variable?

Tabla	Variable	Tipo	Valores permitidos	¿Puede estar vacío?	Descripción	De dónde viene
transacciones	id_transaccion	Texto	Código único	No	Identificador de cada compra	Sisten
transacciones	monto	Número	Mayor a 0	No	Cantidad en pesos	Termi
transacciones	fecha_hora	Fecha	Formato YYYY-MM-DD	No	Cuándo ocurrió	Sisten
transacciones	es_fraude	Sí/No	Verdadero o Falso	No	¿Es fraudulenta?	Analis

Mantén este archivo actualizado y guárdalo como `diccionario_datos.csv`

¿De dónde obtengo mis datos?

¿Qué datos necesito?

- Identifica todas las variables en tu diccionario.
- Asegúrate que cada variable tenga una fuente definida.
- Determina la frecuencia de recolección necesaria.

Métodos de recolección

- **Fuentes directas**: encuestas que yo diseño, experimentos, observaciones.
- **Fuentes existentes**: bases de datos públicas, APIs, archivos institucionales.
- **Herramientas de captura**: formularios digitales, sensores, programas de descarga automática.

Planificación del muestreo

- Define claramente: ¿a quién o qué voy a estudiar?
- ¿Cuántos casos necesito para que sea representativo?
- ¿Hay grupos específicos que debo incluir?

Control de calidad de los datos

Prevención (al momento de capturar)

- Usar formularios que validen automáticamente los datos.
- Crear listas de opciones cerradas cuando sea posible.

Detección de errores (después de capturar)

- Verificar que no hay duplicados donde no debería haberlos.
- Revisar que los valores estén en rangos razonables.
- Comprobar que las fechas sean lógicas.

Documentación

- Registrar de dónde viene cada dato y cuándo se capturó.
- Mantener los datos originales sin modificar en una carpeta separada.

Consideraciones éticas y de privacidad

- **Minimiza la recolección**: solo recolecta los datos que realmente necesitas.
- **Protege la identidad**: usa códigos en lugar de nombres cuando sea posible.
- **Controla el acceso**: define quién puede ver qué información.
- **Datos sintéticos**: si necesitas compartir datos, considera generar versiones artificiales que mantengan las características importantes pero no revelen información personal.

¿Qué son los datos sintéticos?

Los **datos sintéticos** son conjuntos de datos artificiales que imitan las características estadísticas de datos reales sin contener información real de personas específicas.

¿Cuándo usarlos?

- Para compartir datos con colaboradores
- Para publicar datasets de investigación
- Para entrenar modelos sin exponer datos sensibles
- Para crear ejemplos educativos

Ventajas

- Protegen la privacidad individual
- Mantienen patrones estadísticos útiles
- Permiten reproducibilidad
- Eliminan restricciones legales de uso

Ejemplo práctico: Dataset de estudiantes

Datos originales (sensibles):

ID	Nombre	Edad	Promedio	Ciudad
1	María García	22	8.5	Guadalajara
2	Juan López	24	7.2	Monterrey

Datos sintéticos (seguros para compartir):

ID	Edad	Promedio	Ciudad
1	23	8.3	Ciudad_A
2	25	7.0	Ciudad_B

> Los datos sintéticos mantienen las relaciones entre edad y promedio, pero eliminan nombres reales y codifican ubicaciones.

Herramientas básicas:

- **Python**: librerías como `faker` o `synthpop`
- **R**: paquetes como `synthpop` o `simPop`
- **Excel**: funciones aleatorias con distribuciones controladas

Organización de archivos para tu proyecto

```
/datos
  /originales       # datos tal como los obtuviste (no tocar)
  /procesados       # datos limpios listos para análisis
  diccionario_datos.csv
/documentos
  esquema_datos.png # tu diagrama de organización
README.md          # explicación del proyecto
```

Herramientas recomendadas

- **Excel/Google Sheets**: para empezar y datos pequeños.
- **SQLite**: base de datos simple, ideal para aprender.
- **R/Python**: para análisis más avanzados.
- **Google Forms**: para recolectar datos con formularios.

> Elige la herramienta según el tamaño de tus datos y tu experiencia.

Lista de verificación para la Evidencia 4A

- [ ] **Archivo de base de datos** creado (.xlsx, .csv o base de datos simple).
- [ ] **Herramienta utilizada** especificada (Excel, SQLite, etc.).
- [ ] **Variables incluidas** listadas (ejemplo: “20 variables, incluyendo 5 numéricas y 15 categóricas, con N=100 observaciones”).
- [ ] **Diccionario de datos** completo y actualizado.
- [ ] **Video corto** (máximo 3 minutos) mostrando tu base de datos.
- [ ] **Documento PDF** con enlace al video y capturas de pantalla.

Actividad

1. **Dibuja** la organización de tus datos (2-3 grupos principales).
2. **Define** los campos principales de cada grupo (8-10 variables clave).
3. **Escribe** 5 reglas de calidad para tus datos.
4. **Describe** tu plan de recolección (de dónde obtendrás los datos).

Ejemplo práctico - Estructura de datos

```sql
-- Tabla principal de transacciones
CREATE TABLE transacciones (
  id_transaccion TEXT PRIMARY KEY,
  id_cliente TEXT NOT NULL,
  id_tienda TEXT NOT NULL,
  fecha_hora TIMESTAMP NOT NULL,
  monto DECIMAL(10,2) CHECK (monto >= 0),
  canal TEXT CHECK (canal IN ('pos','web','app','telefono')),
  es_fraude BOOLEAN NOT NULL DEFAULT 0
);

-- Tabla de clientes
CREATE TABLE clientes (
  id_cliente TEXT PRIMARY KEY,
  segmento TEXT,
  fecha_registro DATE
);

-- Tabla de tiendas
CREATE TABLE tiendas (
  id_tienda TEXT PRIMARY KEY,
  categoria TEXT,
```

> Este es solo un ejemplo. Adáptalo a tu proyecto específico.

Puntos clave para recordar

- **Planifica antes de recolectar**: diseña tu estructura de datos antes de empezar a capturar información.
- **Documenta todo**: mantén un diccionario actualizado de qué significa cada variable.
- **Piensa en la calidad**: establece reglas para asegurar datos confiables.
- **Considera la ética**: protege la privacidad y recolecta solo lo necesario.

> La calidad de tu análisis depende directamente de la

Recursos adicionales (1/2)

Aprendizaje de SQL y bases de datos

- **SQL Bolt** - Tutorial interactivo de SQL: https://sqlbolt.com/
- **W3Schools SQL** - Tutorial básico de SQL: https://www.w3schools.com/sql/
- **DB Browser for SQLite** - Herramienta visual para SQLite: https://sqlitebrowser.org/
- **Kaggle Learn** - Cursos gratuitos de bases de datos: https://www.kaggle.com/learn

Herramientas para recolección de datos

- **Google Forms** - Para crear formularios de captura: https://forms.google.com/
- **Microsoft Forms** - Alternativa a Google Forms: https://forms.microsoft.com/
- **SurveyMonkey** - Para encuestas más complejas: https://www.surveymonkey.com/

Tip

Para empezar: Comienza con SQL Bolt para aprender consultas básicas y usa Google Forms para crear tus primeros formularios de recolección.

Recursos adicionales (2/2)

Datos sintéticos y privacidad

- **Faker (Python)** - Librería para generar datos sintéticos: https://faker.readthedocs.io/
- **Synthpop (R)** - Paquete para datos sintéticos: https://cran.r-project.org/package=synthpop
- **SDV** - Synthetic Data Vault: https://sdv.dev/
- **Mostly AI** - Plataforma de datos sintéticos: https://mostly.ai/

Ética y buenas prácticas

- **FAIR Data Principles** - Principios para datos FAIR: https://www.go-fair.org/fair-principles/
- **Guía de ética en datos del MIT**: https://ethics.fast.ai/

> **Note**
>
> **Recursos avanzados**: Estos recursos son útiles cuando ya tengas experiencia básica y necesites proteger datos sensibles o