

SUBSECRETARÍA DE POSGRADO · FHyCS · UNJu · JUNIO 2026

UNIDAD 3 · MATERIAL COMPLEMENTARIO DE LECTURA**Construcción y Gestión de Bases de Datos
para Proyectos de Investigación***Dr. Javier Altamirano · Mg. Ana María Chalabe*

Facultad de Humanidades y Ciencias Sociales – UNJu

Introducción

Este material de lectura acompaña el tercer encuentro del curso y tiene como propósito desarrollar en profundidad los conceptos trabajados durante la clase sincrónica. Su lectura previa o posterior a la sesión en vivo le permitirá afianzar los fundamentos de la construcción y gestión de bases de datos, y contar con una referencia escrita para las actividades prácticas de la unidad.

Mientras la Unidad 2 se concentró en el diseño conceptual de la base de datos —qué registrar y por qué—, esta unidad avanza hacia su construcción efectiva: cómo generar datos propios mediante formularios, cómo evaluar y depurar datos que provienen de terceros, y cómo automatizar la integración de múltiples fuentes mediante Power Query.

El material está organizado en seis secciones: fuentes primarias y el formulario como instrumento de recolección; evaluación y depuración de fuentes secundarias; el complemento ExcellInfo Add-in; Power Query como herramienta de integración; criterios de calidad para bases de datos de investigación; y síntesis aplicada a las actividades de la unidad.

UNIDAD 03 — Construcción y Gestión de Bases de Datos para proyectos de investigación científica

1. Fuentes primarias: el formulario como instrumento de recolección

Un formulario digital es la interfaz entre el investigador y el informante. Su función es capturar datos de manera estandarizada, evitando los errores que introduce el ingreso manual posterior a la recolección. En términos de base de datos, cada pregunta del formulario corresponde a un campo, y cada respuesta obtenida genera un nuevo registro.

A diferencia de un cuestionario en papel, el formulario digital ofrece ventajas concretas para la integridad de los datos: obliga al tipo de dato correcto, permite establecer rangos válidos y opciones cerradas, elimina el paso intermedio de transcripción manual, y registra automáticamente metadatos como fecha y hora de respuesta.

1.1 Google Forms: características y flujo de trabajo

Google Forms es una herramienta gratuita de Google que permite crear formularios en línea con integración directa a Google Sheets. No requiere cuenta institucional ni instalación de software, funciona desde cualquier dispositivo con acceso a internet, y las respuestas se vuelcan automáticamente en una hoja de cálculo en tiempo real. La hoja resultante se exporta a formato .xlsx en un solo paso, quedando lista para depurar y analizar.

El flujo de trabajo típico con Google Forms en un proyecto de investigación tiene cinco etapas:

Diseño	Configurar una pregunta por cada campo del diccionario de datos, con el tipo de pregunta correcto.
Distribución	Compartir el enlace del formulario con los informantes.
Revisión	Verificar las respuestas a medida que llegan, en la solapa "Respuestas".
Exportación	Descargar la hoja de Google Sheets vinculada como archivo .xlsx.
Depuración	Revisar y corregir los datos exportados antes de tratarlos como base definitiva.

1.2 El instrumento no siempre es un formulario digital

No todos los proyectos de investigación recolectan datos propios mediante encuestas o formularios estructurados. Una investigación basada en entrevistas en profundidad necesita una guía de entrevista organizada según los ejes temáticos de las variables de análisis, no un formulario cerrado. Una investigación documental o archivística necesita una ficha de relevamiento que sistematice la lectura de fuentes escritas. Y un proyecto que se apoya en fuentes secundarias existentes no necesita recolectar nada nuevo: necesita una evaluación crítica de la fuente que ya tiene disponible.

El criterio que decide qué instrumento usar no es la preferencia personal, sino la coherencia con la metodología del proyecto: el instrumento debe garantizar que cada variable definida en el diccionario de datos tenga un campo o una sección equivalente donde quede registrada de manera sistemática.

La elección del instrumento de recolección es parte del diseño metodológico, no un detalle técnico posterior. Un formulario digital es la opción más eficiente cuando la investigación recolecta datos estructurados de muchos casos, pero no es la única opción válida ni siempre la más adecuada.

2. Fuentes secundarias: evaluación y depuración

Las fuentes secundarias son bases de datos generadas por otras personas u organizaciones para fines distintos al proyecto propio. Como se desarrolló en la Unidad 2, su uso es legítimo y frecuente, pero implica un conjunto de operaciones previas al análisis que no pueden omitirse: evaluar su calidad, comprender su codificación original y depurar los problemas que presenten.

2.1 Problemas frecuentes en datos de terceros

Los problemas más habituales que se encuentran al recibir una base de datos de terceros —especialmente registros administrativos como los que produce el sistema de salud, organismos públicos o instituciones— son los siguientes:

Tipo de problema	Ejemplo concreto
Encabezados con espacios o formato inconsistente	Una columna llamada " ID" con espacios antes del nombre
Columnas vacías o irrelevantes	Columnas sin encabezado ni datos, generadas por exportaciones automáticas
Texto no normalizado	Un campo de sexo con valores: M, F, MASCULINO, Femenino, vacío
Códigos sin documentar	Un campo de situación de caso con valores C, D, 1 — sin diccionario que indique su significado
Formato de fecha u hora inconsistente	Una columna de fecha de hisopado con valores de texto, fechas y números seriales mezclados
Valores faltantes sin criterio	Campos vacíos en columnas clave como sexo o edad, sin código de dato faltante
Tipos de dato incorrectos	Números almacenados como texto, lo que impide hacer cálculos o agrupaciones

Antes de depurar, conviene hacer una evaluación diagnóstica que responda tres preguntas: ¿cuántos registros tiene la base y cuántos son válidos?, ¿qué campos tienen datos faltantes y en qué proporción?, y ¿cuáles son los valores únicos de cada campo categórico? Esta exploración inicial —generalmente con filtros de Excel— evita depurar a ciegas y permite priorizar los problemas más relevantes para el análisis que se quiere realizar.

Toda modificación sobre una base de datos secundaria debe documentarse. Es práctica indispensable conservar siempre la versión original sin modificar y trabajar sobre una copia. Las decisiones de depuración forman parte de la metodología del proyecto y deben poder justificarse.

3. ExcellInfo Add-in: herramientas de depuración

ExcellInfo Add-in es un complemento gratuito para Microsoft Excel que, una vez instalado, aparece como una solapa adicional en la cinta y pone a disposición un conjunto de herramientas diseñadas específicamente para la normalización y limpieza de datos. El complemento y su video tutorial de instalación están disponibles en el Aula Virtual.

Su ventaja principal es que opera directamente sobre los valores de las celdas, sin requerir fórmulas auxiliares. Esto simplifica el flujo de trabajo cuando se trabaja con bases grandes, porque no genera columnas temporales que luego hay que eliminar.

3.1 Herramientas principales

Herramienta	Función
Cambiar mayúsculas/minúsculas	Convierte texto a MAYÚSCULAS, minúsculas o Tipo Oración en un rango seleccionado
Eliminar espacios	Suprime espacios duplicados, iniciales y finales en celdas de texto
Eliminar duplicados avanzado	Detecta y marca o elimina filas duplicadas con opciones configurables
Normalizar texto	Aplica reglas de estandarización sobre campos de texto
Proteger/desproteger hojas	Gestiona la protección de hojas con o sin contraseña
Filtros personalizados	Aplica filtros con criterios avanzados sobre el rango activo

3.2 ExcellInfo Add-in y Power Query: roles complementarios

Es importante distinguir el rol de cada herramienta dentro del flujo de depuración. Power Query se ocupa de la importación y la estructura: detecta delimitadores, elimina columnas vacías, ajusta tipos de dato y permite combinar archivos antes de cargar los datos en Excel. ExcellInfo Add-in, en cambio, opera sobre los valores una vez que los datos ya están cargados en una hoja: normaliza texto, elimina espacios y depura inconsistencias dentro de las celdas.

En la práctica, una base de datos administrativa como un registro de casos de salud suele requerir ambas herramientas en secuencia: primero Power Query para resolver la estructura del archivo de origen (por ejemplo, un CSV con columnas vacías y encabezados mal formateados), y luego ExcellInfo Add-in para resolver las inconsistencias de contenido dentro de los campos ya cargados.

No existe un orden único para depurar una base de datos, pero una secuencia eficiente es: hacer una copia de seguridad, explorar la estructura, resolver los problemas de importación con Power Query, normalizar los valores con ExcellInfo Add-in, y por último verificar los tipos de dato antes de analizar.

4. Power Query: integración y transformación de datos

Power Query es una herramienta integrada en Microsoft Excel (desde la versión 2016 en adelante, y también en Power BI) que permite conectar, importar, transformar y cargar datos desde múltiples fuentes de manera automatizada y reproducible. Su diferencia fundamental respecto a las operaciones manuales es que cada transformación queda registrada como un paso dentro de una consulta: las transformaciones son editables, no modifican el archivo original, y pueden actualizarse automáticamente cuando cambia el dato de origen.

Power Query se accede desde la cinta de Excel a través de la solapa Datos → Obtener y transformar datos. Las fuentes que puede conectar incluyen archivos Excel, archivos de texto y CSV, carpetas completas para procesar múltiples archivos en bloque, páginas web con tablas HTML, y bases de datos SQL, entre otras.

4.1 Importar y depurar un archivo CSV

Los archivos CSV —con coma o punto y coma como separador— son uno de los formatos más comunes para la distribución de datos de registros administrativos. Su apertura directa en Excel produce a veces problemas de codificación, separadores mal detectados o tipos de dato incorrectos. Power Query resuelve esto de manera más controlada que la importación directa:

Conectar	Datos → Obtener datos → Desde archivo → Desde texto/CSV. Power Query detecta automáticamente el delimitador.
Transformar	En el editor: eliminar columnas vacías, corregir encabezados, ajustar tipos de dato por columna.
Cargar	Cerrar y cargar la consulta en una nueva hoja de Excel.
Depurar	Sobre los datos ya cargados, aplicar ExcelInfo Add-in para normalizar texto y valores categóricos.
Analizar	Con la base depurada, construir tablas dinámicas para responder preguntas concretas (por ejemplo, casos por mes).

Este flujo —importar con Power Query, depurar con ExcelInfo Add-in, analizar con tabla dinámica— es el que se aplica en la actividad práctica de esta unidad sobre una base real de casos de Covid-19 del año 2021.

4.2 Consolidar múltiples archivos

Uno de los usos más frecuentes de Power Query en investigación es la consolidación de múltiples archivos con la misma estructura en una sola base de datos. Este escenario aparece cuando los datos se reciben por período, por departamento o por institución. El procedimiento es: guardar todos los archivos en una misma carpeta, conectar con Datos → Obtener datos → Desde archivo → Desde carpeta, seleccionar "Combinar y transformar datos", verificar que los encabezados sean consistentes entre archivos, y cerrar y cargar la base consolidada.

Un ejemplo de este escenario es el Registro Ganadero, donde los datos de marcas y señales de hacienda están distribuidos en 16 archivos separados, uno por departamento de la provincia. Una vez consolidados, una tabla dinámica permite responder preguntas que serían muy laboriosas de calcular archivo por archivo: por ejemplo, la cantidad total de vacunos, ovinos y caprinos registrados por departamento y por año. Para este tipo de análisis conviene ubicar el departamento y el año en el área de Filas (o uno en Filas y otro en Columnas), y la categoría de hacienda como Filtro o como Columnas adicionales, con la función Suma en el área de Valores.

Para que la consolidación funcione correctamente, todos los archivos deben tener exactamente los mismos encabezados y en el mismo orden. Si algún archivo tiene columnas adicionales o en distinto orden, Power Query puede producir resultados incorrectos.

4.3 Importar datos desde la web

Power Query puede conectarse directamente a páginas web que contengan tablas HTML y extraer esos datos como si fueran un archivo local. Esta funcionalidad es útil para incorporar datos de portales estadísticos, organismos públicos o enciclopedias en línea. El procedimiento es: Datos → Obtener datos → Desde otras fuentes → Desde la web, pegar la URL, seleccionar la tabla de interés en el panel de navegación, aplicar las transformaciones necesarias, y cargar.

Una vez que los datos importados están cargados en Excel, suele ser útil representarlos visualmente. Un gráfico de barras o de líneas construido a partir de la tabla importada permite comunicar de forma inmediata una tendencia, una comparación o una distribución que en formato tabular requeriría una lectura más detenida. La elección del tipo de gráfico depende de la naturaleza de los datos: líneas para series temporales, barras para comparar categorías, y gráficos circulares solo cuando las categorías son pocas y se quiere mostrar proporción sobre un total.

Cuando se extrae información de sitios web para un proyecto de investigación, es importante registrar la URL completa y la fecha de acceso, ya que el contenido puede cambiar. Estos datos forman parte de la fuente y deben citarse adecuadamente.

5. Criterios de calidad para bases de datos en investigación

La calidad de una base de datos no se mide únicamente por la ausencia de errores. En el contexto de la investigación científica, una base de datos de calidad debe cumplir con un conjunto de criterios que garanticen su validez, su reproducibilidad y su utilidad para el análisis.

Criterio	Descripción
Compleitud	Los campos críticos no tienen valores faltantes, o el porcentaje está documentado y dentro de lo esperado
Consistencia	Los valores de cada campo pertenecen al conjunto válido definido; no hay variantes no estandarizadas
Unicidad	No existen registros duplicados; cada unidad de análisis aparece una sola vez
Validez	Los valores están dentro de los rangos posibles para cada variable
Trazabilidad	El origen de los datos, las transformaciones aplicadas y las decisiones de depuración están documentados
Reproducibilidad	El proceso de construcción y limpieza de la base puede ser replicado por otro investigador con los mismos resultados

Power Query aporta directamente a la trazabilidad y la reproducibilidad: cada paso de transformación queda registrado y puede revisarse o repetirse. ExcellInfo Add-in aporta a la consistencia, normalizando los valores de texto. Ninguna herramienta por sí sola garantiza la calidad: lo que la garantiza es el criterio metodológico con que se las usa.

Una base de datos perfecta rara vez existe en la práctica de la investigación de campo. Lo que distingue una investigación sólida no es la ausencia total de problemas en los datos, sino la capacidad de identificarlos, documentarlos y tenerlos en cuenta al momento de interpretar los resultados.

6. Síntesis y aplicación a las actividades de la unidad

Esta unidad recorrió tres componentes de la construcción de una base de datos: la generación de datos propios mediante un instrumento de recolección, la evaluación y depuración de fuentes secundarias, y la integración de múltiples fuentes mediante Power Query. Estos tres componentes se aplican de manera directa en las actividades asincrónicas del Trabajo Práctico N° 3.

La primera parte de la actividad práctica retoma el flujo de depuración con una base real de casos de Covid-19 del año 2021: importación con Power Query, corrección de encabezados y columnas vacías, normalización de valores con ExcellInfo Add-in, y análisis final mediante tabla dinámica de casos por mes. La segunda parte aplica Power Query de manera autónoma sobre dos escenarios distintos: la consolidación del Registro Ganadero con un análisis de hacienda por departamento y año, y la importación de una tabla desde la web acompañada de un gráfico.

Fuente primaria	¿Qué instrumento de recolección corresponde a tu proyecto? ¿Está alineado con tu diccionario de datos?
Fuente secundaria	¿Qué problemas de calidad tiene la base que vas a usar? ¿Requiere Power Query, ExcellInfo Add-in, o ambos?
Integración de fuentes	¿Tu proyecto necesita combinar varios archivos o incorporar datos de la web? Power Query automatiza ambos casos.
Documentación	¿Registraste las decisiones de depuración y, si corresponde, la URL y fecha de acceso de fuentes web?

En la Unidad 4 — el encuentro final del curso — se trabajará con las bases de datos de cada participante para realizar análisis exploratorios, construir visualizaciones y explorar el uso de herramientas de inteligencia artificial como apoyo al proceso de investigación. Cuanto más sólida sea la base construida en esta unidad, mayor será el provecho de ese encuentro.

Referencias bibliográficas

Mayo, F. (2024). *Excel para el análisis de datos: Técnicas y trucos para simplificar el trabajo*. Madrid: Anaya Multimedia. 400 pp. ISBN: 978-84-415-5035-3. (Capítulos 4-6)

Microsoft. (s.f.). *Acerca de Power Query en Excel*. Microsoft Learn. <https://learn.microsoft.com/excel/power-query>

Romero-Carazas, R., Mayta-Huiza, D., Ancaya-Martínez, M., Tasayco-Barrios, S. y Berrio-Quispe, M. (2024). *Método de investigación científica: Diseño de proyectos y elaboración de protocolos en las Ciencias Sociales*. Lima: Editorial Idicap Pacífico. 198 pp. <https://doi.org/10.53595/eip.012.2024> (Capítulo 4)