

Informe Práctica 2024

15 MARZO

Oficina de Estudios y Estadísticas
División de Políticas Públicas
Subsecretaría de Ciencia, Tecnología,
Conocimiento e Innovación

José Guillermo Sepúlveda Salazar





Contenido

Introducción	3
Objetivo general	3
Objetivos específicos	3
Necesidades de la Institución	4
Recomendaciones	4
1. Análisis de indicadores	4
2. Mantenimiento de datos OCDE	5
3. Observaciones tablas de indicadores I+D	6
Propuestas	7
1. Etiquetado de Tablas	7
2. Modelo Entidad-Relación para la Encuesta sobre Gasto y Personal en Investigación y Desarrollo (I+D) y para los datos extraídos desde el Sistema de Información de Educación Superior (SIES)	8
Justificación de la Propuesta	10
Conclusión.....	12
Anexos	13
Anexo 1: Modelo de Extracción, Transformación y Carga (ETL, por sus siglas en inglés). .	13



Introducción

Se establece un marco para fortalecer los procesos estadísticos de la Oficina de Estudios y Estadística (OEE), particularmente en la Unidad de Estadísticas, el enfoque se centra en la actualización y optimización de los procesos de recopilación y análisis de datos, asegurando la entrega de información precisa y relevante. Este plan contempla el uso de tecnologías actualizadas y metodologías eficientes y seguras para la gestión de datos, para la toma de decisiones informadas dentro del sector.

Objetivo general

“Contribuir al fortalecimiento de los productos elaborados por la Unidad de Estadística, con el fin de proporcionar a la Oficina y al Ministerio datos confiables y actualizados del sistema CTCI.”

Objetivos específicos

- Objetivo específico 1: Actualizar indicadores de encuestas y datos administrativos asociados a CTCI en la plataforma Observa.
- Objetivo específico 2: Diseñar y proponer un diagrama de Entidad-Relación¹ detallado para I+D (Investigación y Desarrollo) a la Oficina de Estudio y Estadística.
- Objetivo específico 3: Implementar un sistema automatizado de extracción de información en la web a indicadores establecidos.

La estructura de este documento se organiza en varias partes claves: primero las “Necesidades de la Institución”, donde se presentan los hallazgos principales a modo de diagnóstico, luego las “Recomendaciones” en las cuales se presentan los productos realizados que mejoran los procesos de análisis estadísticos, en el apartado de “Propuestas” se introducen innovaciones en la estructura de datos, las cuales incluyen un etiquetado de tablas con un formato definido y variables estandarizadas que las componen. La “Conclusión” sintetiza los diagnósticos y recomendaciones, destacando su importancia en la consolidación estadística. Finalmente, en los “Anexos”, se ilustra un flujo de trabajo que implementará la unidad de datos, lo cual se relaciona con los objetivos de la práctica.

¹¿qué es Entidad-Relación?: <https://universidadeuropea.com/blog/modelo-entidad-relacion/>



Necesidades de la Institución

La principal labor indicada al practicante es la actualización de los resultados de las Encuesta sobre Gasto y Personal en Investigación y Desarrollo (I+D) año 2021, y de los datos extraídos desde el Sistema de Información de Educación Superior (SIES) año 2022 y 2023. Donde se ha detectado la posibilidad de optimizar las tablas de la base de datos, las cuales presentan cierto nivel de desvinculación y duplicidad. Esta situación, agravada por la estructura actual basada en SQL (Structured Query Language) que establece un modelo de datos relacionales, carece de una relación entre sus tablas, lo que podría generar costos elevados a largo plazo, especialmente al almacenar más datos. En este contexto, se busca sentar las bases para el desarrollo de un sistema más robusto en el futuro.

Por otro lado, se han reconocido procesos que requieren una inversión de tiempo y esfuerzo manual para localizar datos relevantes, debido a la falta de documentación adecuada, lo que conducen a retrasos y obstáculos en el desarrollo. Para solucionar esto, se propone mejorar la eficiencia en la extracción de datos de la web a través de la automatización. Aunque las metodologías para lograrlo no son excesivamente complejas, se espera que resulten en una mejora sustancial en la eficiencia de la actualización de datos.

El desarrollo de las aplicaciones se realizará utilizando Python versión 3.9² como el lenguaje de programación principal y Markdown³ para la creación de documentación, todo dentro del entorno de archivos Jupyter Notebook. Para la gestión de librerías se implementará un entorno virtual utilizando Miniconda⁴.

Recomendaciones

A continuación, se presentan los productos que derivan del trabajo de práctica durante este periodo. Para mayor facilidad en su lectura, estas recomendaciones son separadas por las principales actividades realizadas.


1. Análisis de indicadores

- Los documentos Jupyter presentan un método sistemático para la automatización de indicadores, abarcando desde la recolección de datos en línea hasta su procesamiento y almacenamiento en Sistemas de Gestión de Bases de Datos (SGBD). Se establece un marco metodológico con sólidas validaciones para la extracción de archivos, asegurando su adecuada integración en el repositorio correspondiente. Adicionalmente, se verifica dentro del SGBD que los indicadores estén actualizados al año correspondiente.
- Para fortalecer la seguridad y garantizar la privacidad de los datos, el sistema incluye mecanismos de autenticación. Esta estrategia mejora no solo la eficiencia en la gestión de indicadores, sino que también preserva la confidencialidad de la información, alineándose con los estándares de seguridad para el tratamiento de datos sensibles.

²¿qué es Python?: <https://docs.python.org/es/3/tutorial/>

³¿qué es Markdown?: https://markdown.es/#google_vignette

⁴¿qué es Miniconda?: <https://docs.anaconda.com/free/miniconda/index.html>

- 
- El proceso automatizado se divide en tres etapas clave: extracción, transformación y carga (ETC). Inicialmente, el sistema verifica la disponibilidad de la base de datos en el repositorio. Si no se encuentra, se procede a extraer los datos mediante técnicas de Web Scraping o conexiones API, incluyendo la descarga y el nombramiento adecuado de los archivos. Durante la transformación, se analizan estadísticamente los indicadores para obtener resultados actualizados. Finalmente, en la fase de carga, el sistema se conecta al SGBD con credenciales específicas y confirma si los indicadores están al día. Si no lo están, se cargan los datos transformados.
 - Para mejorar la robustez del sistema y evitar la carga accidental de datos no verificados o erróneos en el SGBD, se han incorporado mecanismos de prevención en la arquitectura del programa. Antes de proceder a la fase final de carga de los datos transformados, el sistema implementa una capa adicional de validación en algunos indicadores que requiere una intervención manual que consiste en modificar la variable compuesta por "_ensayo" a "_carga". Esta precaución es particularmente crucial dado el volumen significativo de registros manejados en Sistemas de Información de Estadísticas Educativas (SIES), donde la precisión y la integridad de los datos son fundamentales.
 - Este enfoque asegura que solo después de una revisión exhaustiva y la confirmación de la precisión de los resultados transformados, un administrador o usuario autorizado pueda ajustar la variable del sistema para permitir la carga final de los datos. Esta etapa de confirmación actúa como un interruptor de seguridad, asegurando que ningún conjunto de datos se mueva a la etapa de "carga" sin la debida diligencia. Este proceso no solo minimiza el riesgo de corromper la base de datos con información inexacta, sino que también refuerza las prácticas de gobernanza de datos al requerir una validación explícita antes de la integración de nuevos datos, manteniendo así la confiabilidad y la seguridad del sistema en su conjunto.

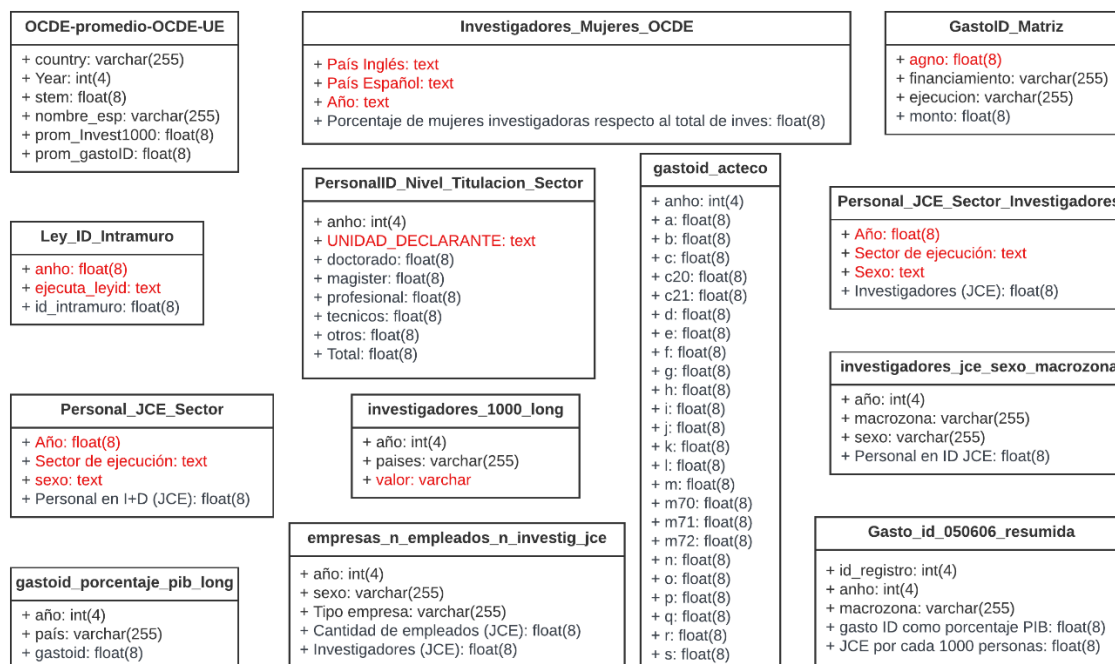
2. Mantenimiento de datos OCDE

- El programa está diseñado para realizar la extracción de datos desde la Organización para la Cooperación y el Desarrollo Económicos (OCDE) mediante la implementación de una interfaz de programación de aplicaciones (API). Posteriormente, ejecuta una consulta sobre tres tablas específicas alojadas en el Sistema de Gestión de Bases de Datos (SGBD): "investigadores_1000_long", "gastoid_porcentaje_pib_long" y "Investigadores_Mujeres_OCDE". La metodología adoptada para el análisis implica la comparación de los resultados obtenidos de estas tablas mediante el uso de estructuras de datos tipo DataFrame.
- En el caso de identificarse diferencias significativas entre los conjuntos de datos (estableciendo previamente un umbral de sensibilidad para la significancia), el programa está capacitado para categorizar y registrar estas discrepancias bajo tres etiquetas distintas: "Diferencia significativa", "Valor no ingresado al SGBD" y "Valor OCDE no reportado". Esta clasificación facilita una interpretación detallada y estructurada de los resultados del análisis.
- Adicionalmente, el programa está configurado para exportar los resultados a un archivo de Excel, incorporando un sistema de versionamiento en el nombre del archivo basado en la fecha y hora de ejecución del programa, lo cual contribuye a la trazabilidad y organización de los archivos generados. Finalmente, se presenta en consola un resumen cuantitativo de los totales correspondientes a cada una de las categorías previamente mencionadas, permitiendo así una rápida visualización del análisis realizado.

3. Observaciones tablas de indicadores I+D

- En el desarrollo del esquema de la base de datos 'ciencia_db', se ha procedido a una meticulosa revisión y diagramación de las estructuras de tablas existentes, donde la representación visual de las tablas ilustra con precisión la nomenclatura como los tipos de datos utilizados para cada variable, tal como se presenta en la *Figura 1*.
- No obstante, un análisis detallado ha revelado ciertas inconsistencias en la nomenclatura de las variables (indicadas en rojo en la *Figura 1*) y tipos de datos que podrían comprometer la integridad y la consistencia de los datos. Además, se ha observado que los nombres de varias variables comparten una similitud notable, lo que sugiere la necesidad de una estandarización para fortalecer la correlación de los datos entre las diferentes tablas y minimizar posibles confusiones.
- Por otro lado, se ha identificado que la denominación actual de las tablas no favorece una interpretación intuitiva de su contenido. Específicamente, algunos nombres resultan excesivamente extensos, desafiando las mejores prácticas de eficiencia en consultas y estructuración de datos. Esta situación subraya la importancia de adoptar un esquema de nomenclatura más lógico y categorizado que promueva una recuperación ágil y eficaz de la información.
- Para abordar estos asuntos, se propone una iniciativa de normalización de variables y un rediseño de la convención de nombres de las tablas, alineando el esquema de la base de datos con estándares de Data Science que faciliten su escalabilidad y mantenimiento a largo plazo. Este proceso no solo mejorará la gestión actual de datos, sino que también sentará las bases para una integración de datos más robusta y coherente en futuras expansiones del proyecto.

Figura 1 Arquitectura original de tablas asociadas a la Encuesta sobre Gasto y Personal en Investigación y Desarrollo.



Fuente: Elaboración propia con base a la información extraída del Sistema de Base de Datos (SGBD) de la Oficina de Estudios y Estadística.

Propuestas

En este capítulo se delinean estrategias claves para la reforma y mejoramiento de la actual base de datos, sugiriendo una normalización y un esquema de nomenclatura clarificada para las tablas que mejorará la gestión y el análisis de datos, proponiendo un diseño Entidad-Relación que optimiza el almacenamiento de estos y facilita las consultas analíticas. Además, se recomienda la estandarización de variables para garantizar la consistencia y precisión de la información. Estas iniciativas están dirigidas a fortalecer la integridad y la eficiencia de la base de datos a largo plazo.

1. Etiquetado de Tablas


La revisión de la estructura de datos para 'observa' ha llevado a una serie de recomendaciones enfocadas en la optimización y claridad en la denominación de las entidades de la base de datos. Se propone una nueva nomenclatura con el formato 'Indicador_Criterio_CaracterísticaRelevante', que mejora la semántica y sistematización de las tablas.

A continuación, en el *Cuadro 1*, se detalla la propuesta de reestructuración:

Cuadro 1: Propuesta de etiquetado por cada tabla de indicadores de la Encuesta sobre Gasto y Personal en Investigación y Desarrollo.

Tabla Original	Propuesta de Tabla	Nombre del Indicador
OCDE-promedio-OCDE-UE	Imasd_Promedio_OCDE_UE	Promedio OCDE y Unión Europea.
investigadores_1000_long	Imasd_InvestCadaMil_OCDE	Cantidad de investigadores(as) cada mil personas trabajando en países de la OCDE.
gastoid_porcentaje_pib_long	Imasd_Gasto_PIB_OCDE	Gasto en I+D respecto al PIB en países de la OCDE.
GastoID_Matriz	Imasd_Gasto_Fuente_Sector	Distribución del gasto en I+D según fuente de financiamiento y sector de ejecución.
gastoid_acteco	Imasd_Gasto_Sector_CIIU	Gasto en I+D por empresas según su sector económico (clasificación CIIU rev.4)
Gasto_id_050606_resumida	Imasd_Gasto_PIB_InvestCadaMil	Gasto en I+D como porcentaje del PIB e investigadores/as cada 1000 personas trabajando por macrozona.
PersonalID_Nivel_Titulacion_Sector	Imasd_Person_Nivel_Sector	Distribución del personal en I+D según nivel educacional y sector de ejecución.
Ley_ID_Intramuro	Imasd_EjecutaLey_Intramuro	Porcentaje de empresas que ejecuta montos bajo la Ley I+D entre las que realizan I+D intramuro.
Investigadores_Mujeres_OCDE	Imasd_InvestMujer_OCDE	Porcentaje de investigadoras mujeres en países de la OCDE.
investigadores_jcesexo_macrozona	Imasd_PersonMujer_Macrozona	Porcentaje del personal dedicado a I+D que es mujer respecto al total de cada macrozona.
Personal_JCE_Sector	Imasd_Person_Sector	Distribución del personal dedicado en I+D según su sexo por cada sector de ejecución.
Personal_JCE_Sector_Investigadores	Imasd_Invest_Sector_JCE	Distribución por sexo de las jornadas completas equivalentes (JCE) trabajadas en I+D por investigadores según sector de ejecución.
empresas_n_empleados_n_investig_jce	Imasd_InvestPersonal_JCE	Porcentaje de investigadores(as) que son mujeres en empresas beneficiadas por la Ley I+D. y Porcentaje de personal en I+D que son mujeres en empresas beneficiadas por la Ley I+D.

Fuente: Elaboración propia.



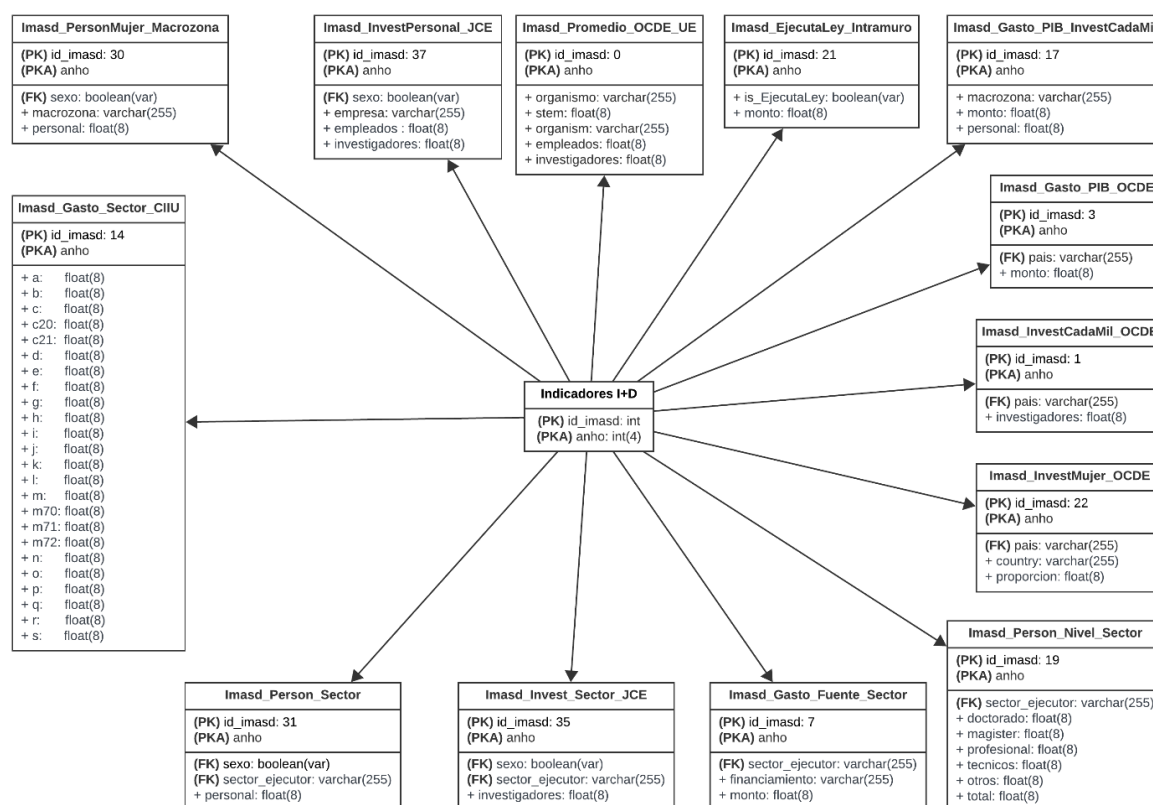
2. Modelo Entidad-Relación para la Encuesta sobre Gasto y Personal en Investigación y Desarrollo (I+D) y para los datos extraídos desde el Sistema de Información de Educación Superior (SIES)

- Tanto la *Figura 2* y la *Figura 3* presentan los diagramas con una versión actualizada de la estructura de la base de datos “ciencia_db” de los indicadores, concebida tras un análisis minucioso del diseño previo. La arquitectura de Modelo Estrella⁵ que se manifiesta es particularmente adecuada para entornos de almacenamiento de datos analíticos, como los Data Warehouse⁶, gracias a su configuración especializada que favorece un rendimiento superior en consultas de informes y análisis de datos.
- Las claves primarias (PK) y compuestas (PKA) se utilizan para identificar de manera única cada registro dentro de las tablas, mientras que las claves foráneas (FK) establecen relaciones entre ellas, permitiendo asociaciones entre distintos conjuntos de datos de manera lógica y coherente. La normalización de variables comunes garantiza que los datos sean consistentes y reduce la redundancia. Esto es crucial para mantener la integridad de los datos a lo largo del tiempo y simplificar las operaciones de mantenimiento y actualización.
- La adopción de una nueva nomenclatura para el nombramiento de tablas con formato “Indicador_Criterio_CaracterísticaRelevante” mejora la legibilidad y la autodescripción de las tablas, lo que promueve una integración fluida y una búsqueda eficiente, factores esenciales para analizar datos de alto nivel.

⁵¿qué es Modelo Estrella? : <https://www.tecon.es/que-es-el-modelo-estrella/#:~:text=El%20Modelo%20Estrella%20facilita%20el,niveles%20de%20detalle%20o%20granularidad.>

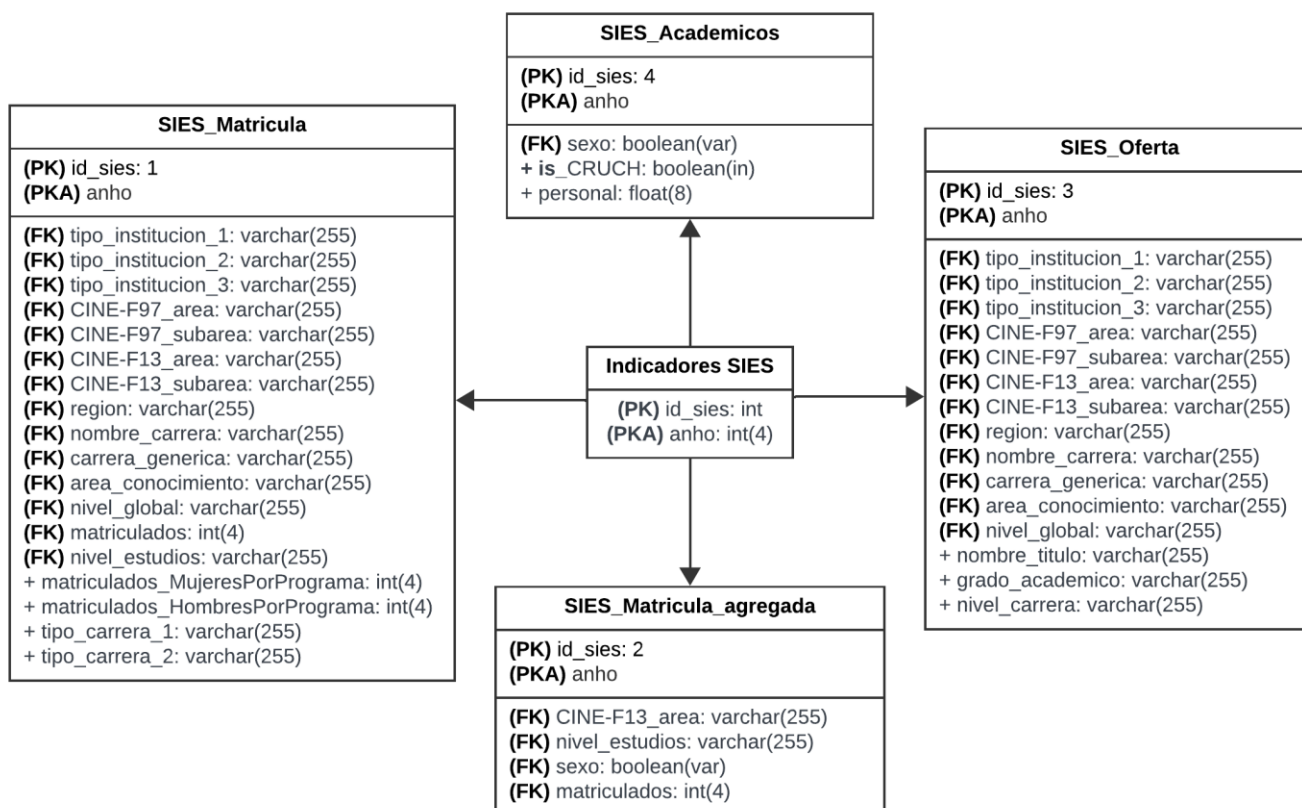
⁶¿qué es Data Warehouse? : <https://www.oracle.com/cl/database/what-is-a-data-warehouse/>

Figura 2: Propuesta de arquitectura de tablas para la Encuesta sobre Gasto y Personal en Investigación y Desarrollo (I+D)



Fuente: Elaboración propia.

Figura 3: Propuesta de arquitectura de tablas para datos extraídos desde el Sistema de Información de Educación Superior (SIES)



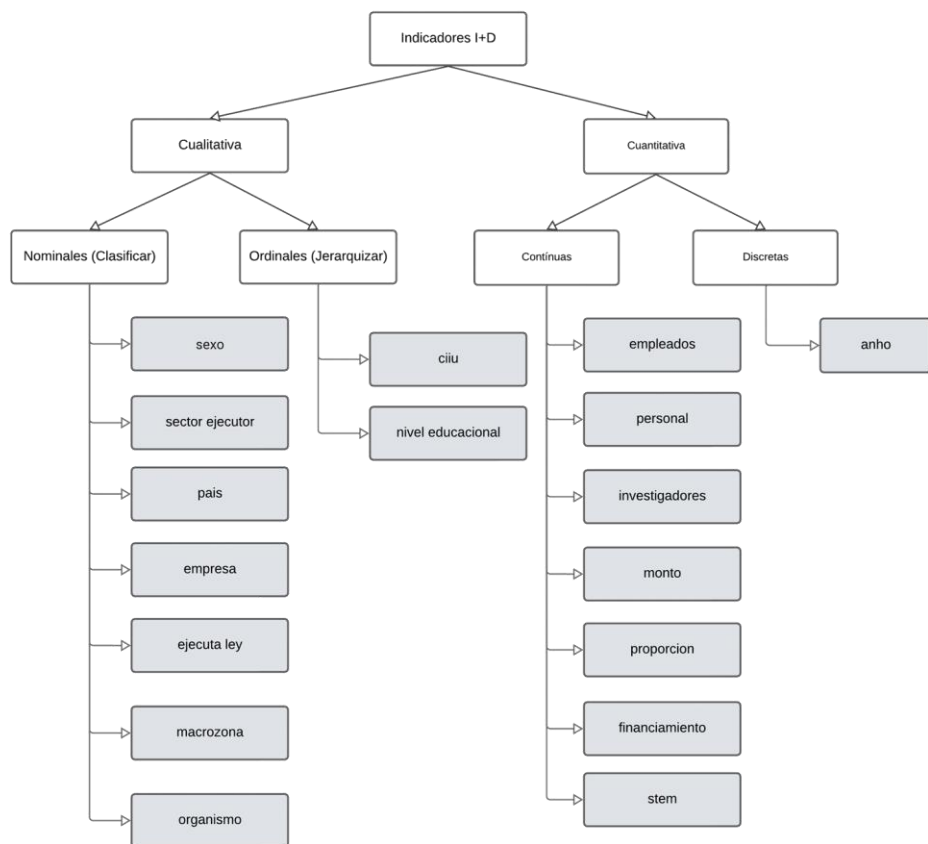
Fuente: Elaboración propia.



Justificación de la Propuesta

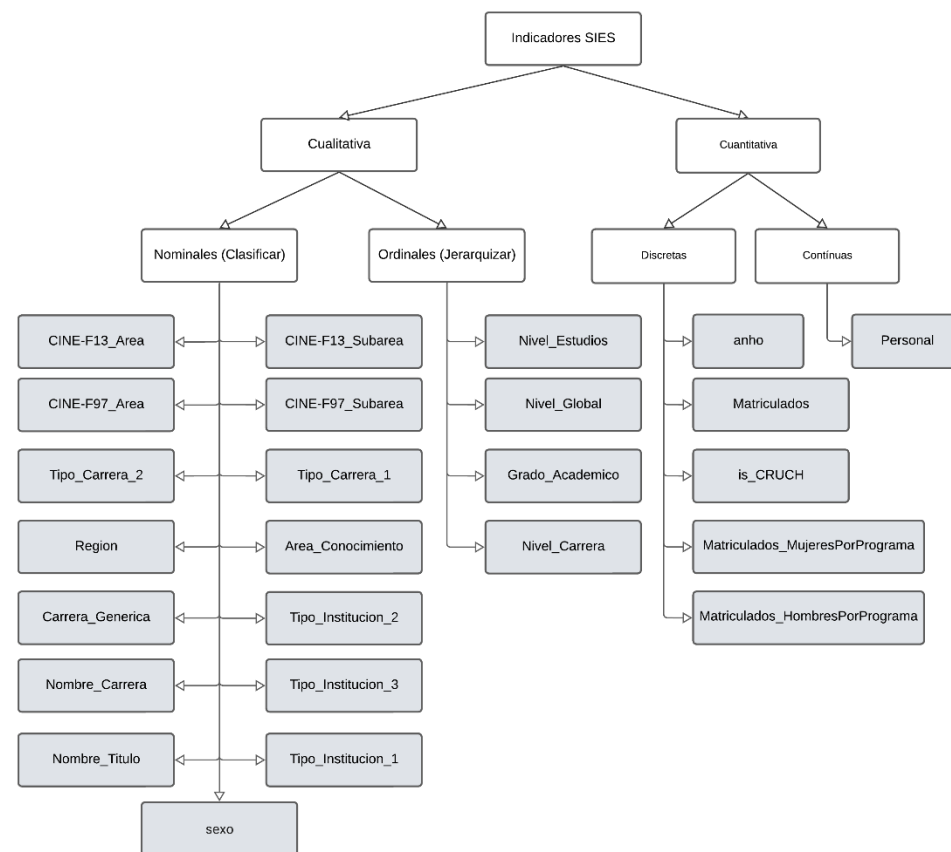
- En la *Figura 4* y *Figura 5* están los diagramas que muestran un enfoque sistemático en la categorización y estandarización de las variables de la base de datos “ciencia_db” de los indicadores de la Encuesta sobre Gasto y Personal en I+D y los datos extraídos del SIES.
- Este esquema elaborado posibilita la interacción entre diversas entidades, facilitando la configuración de un diagrama Entidad-Relación con un diseño Modelo Estrella coherente y funcional. Al identificar y analizar las categorías relevantes, el diagrama garantiza una interconexión lógica de los indicadores, mejorando así la integridad y precisión en la gestión de los datos. Este proceso de diseño promueve una recuperación y análisis de datos eficiente y estructurado, vital para un almacenamiento de datos coherente y accesible para análisis posteriores.

Figura 4: Clasificación por tipo de variables estandarizadas para la construcción de los indicadores de la Encuesta sobre Gasto y Personal en Investigación y Desarrollo (I+D)



Fuente: Elaboración propia.

Figura 5: Clasificación por tipo de variables estandarizadas para los datos extraídos desde el Sistema de Información de Educación Superior (SIES)



Fuente: Elaboración propia.



Conclusión

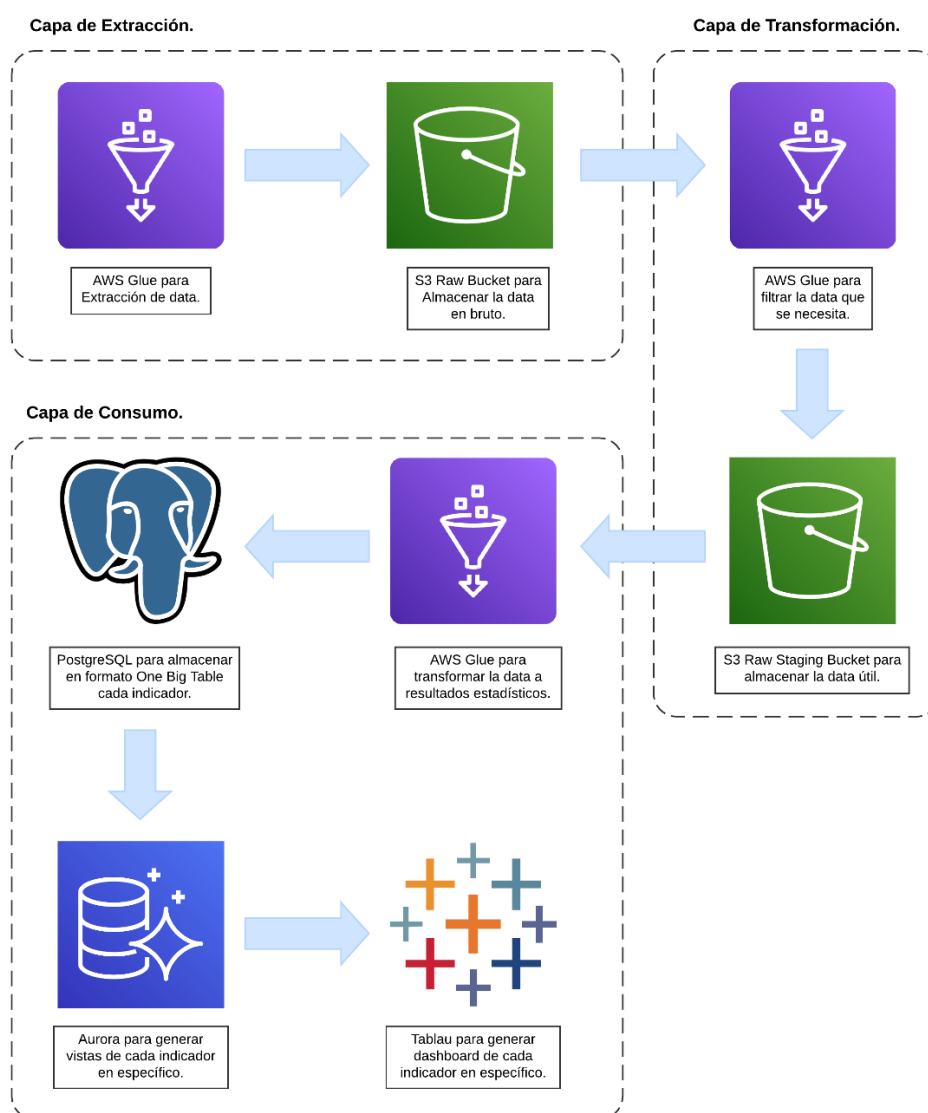
- Este informe culmina con una reflexión sobre los esfuerzos realizados y los caminos a seguir en la mejora de los procesos estadísticos y la gestión de datos de la institución. Habiendo identificado áreas críticas necesitadas de optimización, como la actualización eficaz de los datos de encuestas y la automatización en la recopilación de datos, se han propuesto soluciones innovadoras. Estas recomendaciones no solo abordan la necesidad de una automatización avanzada y un mantenimiento riguroso de datos internacionales, sino que también enfatizan la importancia de una estructuración y estandarización cuidadosa de las bases de datos.
- La adopción de un nuevo esquema de nomenclatura y la implementación de un diseño de Entidad-Relación mejorado son pasos fundamentales hacia la armonización de los datos y la eficiencia operativa. Estas propuestas de reforma tienen el potencial de transformar la gestión de datos, facilitando un análisis más detallado y una toma de decisiones más informada.
- Al mirar hacia el futuro, este informe no solo aboga por soluciones a corto plazo a los desafíos existentes, sino que también establece un marco para el desarrollo sostenible de los sistemas de datos. La visión es clara: avanzar hacia una infraestructura de datos que no solo sea robusta y sistemáticamente organizada, sino que también sea adaptable y escalable frente a las demandas futuras. La implementación exitosa de estas recomendaciones marcará un hito significativo en la evolución de la gestión de datos de la institución, asegurando su capacidad para liderar con conocimiento e innovación en un entorno en constante cambio.

Anexos

Anexo 1: Modelo de Extracción, Transformación y Carga (ETL, por sus siglas en inglés).

- La *Figura 6* ilustra un flujo de trabajo a implementar en la unidad de datos, que comienza con la extracción de datos y culmina en visualizaciones gráficas para la plataforma Observa.
- Incluye el uso integrado de tecnologías AWS y software de terceros, cooperando para producir información precisa y actualizada.
- A partir de la extracción, los datos se someten a un proceso de limpieza y almacenamiento antes de su transformación.
- Luego, estos datos transformados se combinan con otros conjuntos para generar dashboard.
- Este flujo asegura que las representaciones gráficas finales sean tanto confiables como relevantes para el usuario final.

Figura 6 Diagrama de Flujo del Modelo ETL



Fuente: Elaboración propia.