Consolidación y validación de la base ESD/GCE

Documento técnico de ERCE 2025

Juan Carlos Castillo, Daniel Miranda y Katherine Aravena

Santiago de Chile

16 de marzo de 2026

Tabla de contenidos

1 Resumen ejecutivo

Durante la primera etapa del proyecto ESD/GCE ERCE 2025 se desarrolló un proceso de consolidación, estandarización y validación de bases de datos orientado a transformar un conjunto heterogéneo de archivos curriculares en una base analítica única, trazable, reproducible y apta para las siguientes fases del análisis cuantitativo. Esta etapa construye una infraestructura de datos confiable sobre la cual esos análisis puedan realizarse posteriormente con validez metodológica.

El punto de partida estaba compuesto por once módulos temáticos, organizados en distintos archivos y hojas de cálculo, con diferencias de estructura, nomenclaturas no homogéneas, vacíos derivados de celdas combinadas, duplicados operativos, filas separadoras o sin contenido analítico y codificaciones de indicadores que no eran directamente comparables entre sí. En otras palabras, la información necesaria para el análisis ya existía, pero no estaba en una forma que permitiera utilizarla sin riesgo. Analizar ese material tal como se encontraba habría expuesto el proceso a errores de conteo, pérdida de trazabilidad entre registros, confusión entre columnas equivalentes y comparaciones inválidas entre módulos.

La lógica general del trabajo desarrollado en esta entrega fue, por tanto, ordenar sin alterar el contenido sustantivo. El proceso consideró el congelamiento de insumos y la validación del inventario de fuentes; luego, la construcción de mappings por módulo para que estructuras originalmente heterogéneas pudieran traducirse a una estructura canónica común. A partir de ello, se armó una base compartida, se corrigieron artefactos heredados del formato de origen, se separó la información según su función en variables núcleo, indicadores y notas, se recodificaron los indicadores a un dominio comparable, se eliminaron duplicados de manera trazable, se consolidó todo en una base maestra y, finalmente, se ejecutaron controles de calidad en cada etapa. En términos simples, se trató de fijar y revisar los archivos de origen, definir reglas comunes para que todos los módulos hablaran el mismo idioma, ordenar y depurar la información, unirla en una so-

la base y verificar paso a paso que el resultado final fuera consistente, trazable y analíticamente confiable. Cada uno de esos pasos dejó evidencia verificable en forma de archivos intermedios, reportes de control de calidad y decisiones metodológicas documentadas.

El resultado final fue una base master de 7.410 filas y una tabla notes también de 7.410 filas. La base master constituye el insumo principal para el análisis cuantitativo, mientras que la tabla notes se mantuvo como una estructura auxiliar alineada registro a registro. Esta correspondencia uno a uno asegura que, cuando existe información textual o complementaria asociada a una observación, ésta pueda vincularse sin desalineación con la fila analítica correspondiente. No obstante, dado que notes no contiene contenido sustantivo en todos los casos, su valor en esta etapa radica principalmente en la conservación ordenada de esa estructura paralela, más que en aportar evidencia cualitativa completa para cada registro.

La base consolidada tiene una unidad de análisis explícita: cada fila corresponde a una evidencia o cita curricular específica ubicada en un documento, módulo y página determinados, organizando la evidencia curricular bajo una estructura común. Sobre esa unidad de análisis se articulan variables núcleo, variables de trazabilidad e indicadores recodificados.

A ello se suma que la base final no presenta duplicados globales ni de row_uid ni de row_id, que no registra faltantes globales en los campos esenciales module, pais, documento y cita, y que los indicadores quedaron contenidos dentro del dominio esperado, sin valores inválidos fuera de {0,1,NA}. En materia de completitud, se identificaron excepciones puntuales y documentadas en el campo pagina, incluyendo un caso en PAZ y otros casos observados en ESI y Salud. Estos registros no fueron eliminados de la base en esta etapa, sino que se mantuvieron temporalmente por razones de trazabilidad, a la espera de que la información pueda ser corregida o completada si se logra recuperar la referencia correspondiente. Asimismo, el proceso permitió detectar y tratar explícitamente otros casos excepcionales, como tres residuos no binarios en EDS que se conservaron como NA por honestidad semántica.

En síntesis, al comienzo existía la misma información que al final, pero dispersa, heterogénea y en una forma que no permitía confiar plenamente en su uso analítico. Al cierre de la entrega M1, esa misma información quedó organizada, alineada, documentada y verificada. La importancia de esta transformación radica en haber dejado en una forma segura los datos para producir análisis regionales y nacionales defendibles en las fases siguientes del proyecto.

2 Documento técnico de la base ERCE 2025

2.1. Introducción

Este documento corresponde al entregable del Mes 1 de la consultoría asociada al análisis cuantitativo del estudio ESD/GCE ERCE 2025. Su propósito es presentar de manera clara, completa y defendible la metodología utilizada para consolidar y validar la base de datos que servirá como soporte de las etapas analíticas posteriores. En este sentido, la entrega M1 cumple una función fundacional dentro del proyecto: antes de interpretar resultados, comparar países o construir indicadores agregados, era necesario garantizar que el corpus estuviera ordenado, que las unidades de análisis fueran consistentes y que las reglas de procesamiento estuvieran explícitamente documentadas.

La relevancia de esta etapa deriva de una premisa básica del trabajo cuantitativo: los análisis sólo son tan confiables como la base sobre la cual descansan. Cuando los datos provienen de múltiples archivos, hojas, módulos y convenciones de codificación, la producción de una base consolidada deja de ser una tarea secundaria y pasa a ser una condición de posibilidad del análisis mismo. En ese contexto, el objetivo de esta entrega no fue generar conclusiones sustantivas sobre la presencia o distribución de contenidos ESD/GCE, sino construir la infraestructura metodológica que permitirá producirlas de manera válida en los meses siguientes.

En coherencia con los términos de referencia de la consultoría, esta entrega incluye cuatro componentes centrales: la metodología de consolidación de bases de datos, la documentación de decisiones de curatoría y procesamiento, el libro de códigos de la base final y un conjunto de descriptivos univariados que permiten caracterizar el resultado de la consolidación. El valor de este documento reside, por tanto, en explicar no sólo qué producto final existe, sino también por qué puede considerarse confiable, cómo fue construido y qué controles permiten sostener esa afirmación.

2.2. Antecedentes y problema a resolver

El punto de partida del trabajo fue un corpus compuesto por once módulos temáticos: DDHH, ECM, EDS, ESI, Género, HSE, HSE Respeto, PAZ, Pilares actitudinal, Pilares cognitivo y Salud. Estos módulos estaban asociados a distintos archivos fuente y, en algunos casos, a distintas hojas dentro de un mismo libro. El snapshot de referencia utilizado para el run full correspondió al corte 20260225_1928, lo que permitió trabajar sobre un universo estable de insumos y fijar un punto común para todo el proceso posterior.

La heterogeneidad del corpus no era un problema meramente estético o de formato. En la práctica, implicaba que columnas conceptualmente equivalentes podían aparecer con nombres distintos, que un mismo libro podía contener más de una hoja relevante, que algunas cabeceras estuvieran duplicadas o parcialmente vacías, y que ciertos vacíos obedecieran no a ausencia real de información, sino a la lógica visual de celdas combinadas en los archivos de origen. Además, algunos módulos compartían un mismo archivo fuente, mientras que otros requerían inferencias controladas para identificar adecuadamente la hoja o el segmento relevante.

En un nivel más concreto, el problema inicial puede describirse como un desajuste entre el modo en que la información estaba almacenada y el modo en que debía estar organizada para poder ser analizada. Los archivos de origen estaban pensados para contener información curricular y de codificación, pero no necesariamente para operar como una base consolidada de análisis cuantitativo. Esa diferencia es decisiva. Un archivo puede ser útil para lectura humana y, al mismo tiempo, ser riesgoso para el procesamiento si contiene ambigüedades estructurales, vacíos heredados del formato visual, duplicaciones no resueltas o registros cuya señal codificada no puede vincularse de forma confiable con una evidencia textual legible.

Desde el punto de vista analítico, mantener los insumos tal como estaban implicaba varios riesgos. El primero era el riesgo de sobreconteo o subconteo: si existían duplicados no resueltos o filas separadoras tratadas como observaciones reales, los descriptivos y frecuencias podían distorsionarse. El segundo era el riesgo de pérdida de trazabilidad: si no existía una llave operativa estable para seguir cada registro a través de las distintas transformaciones, se debilitaba la posibilidad de auditar el proceso completo. El tercero era el riesgo de comparación inválida: si los indicadores no compartían un dominio común o si columnas análogas estaban definidas de forma distinta entre módulos, cualquier intento de producir síntesis comparables podía volverse

metodológicamente frágil. Finalmente, existía un riesgo de opacidad: incluso si el procesamiento lograba "funcionar", sin documentación explícita de las decisiones sería difícil defender por qué determinadas reglas fueron aplicadas y otras descartadas.

Por estas razones, la primera necesidad del proyecto era ordenar el soporte empírico. Antes de responder preguntas sustantivas, había que resolver una pregunta más básica: cómo transformar un conjunto de insumos heterogéneos en una base única que preservara la información original, pero eliminara las ambigüedades estructurales que impedían su uso analítico directo.

2.3. Enfoque metodológico general

El principio rector del trabajo fue ordenar sin alterar el contenido sustantivo. Esto significa que la consolidación no fue entendida como una intervención destinada a modificar la información, sino como un proceso de estandarización, explicitación de reglas y validación de consistencia. En vez de producir nueva información, el pipeline fue diseñado para asegurar que la información existente quedara correctamente alineada, trazada y preservada en una estructura estable.

Este principio tuvo varias consecuencias metodológicas. En primer lugar, todas las decisiones de procesamiento debían ser explícitas y reproducibles. No bastaba con corregir; era necesario documentar qué problema se observó, qué alternativas se consideraron, qué regla se adoptó y en qué archivos o reportes quedaba trazada esa decisión. En segundo lugar, cada paso del pipeline debía dejar evidencia verificable. Por ello, junto con los archivos principales se generaron salidas de control de calidad, resúmenes de integridad y auditorías específicas por etapa. En tercer lugar, ante casos ambiguos, se privilegió la honestidad semántica por sobre la aparente completitud. Esto fue especialmente importante en la recodificación de indicadores y en el tratamiento de excepciones documentales: cuando un valor no podía interpretarse defensiblemente como 0 o 1, o cuando la evidencia textual asociada a un registro no era legible ni verificable, se prefirió conservar la ambigüedad como NA o mantener el caso en revisión antes que forzar una clasificación engañosa.

En términos simples, el enfoque metodológico puede resumirse así: había que dejar la misma información en una forma mejor estructurada para ser analizada, auditada y protegida frente

a errores de procesamiento. Esa es la base sobre la cual una etapa posterior puede producir resultados sustantivos confiables.

2.4. Metodología de consolidación y validación paso a paso

La consolidación de la base se desarrolló mediante una secuencia de etapas encadenadas, cada una orientada a resolver un problema específico. El valor del pipeline no estuvo sólo en su orden técnico, sino en que cada paso respondió a una necesidad analítica concreta y dejó una huella verificable.

La primera operación consistió en trabajar sobre un conjunto congelado de insumos. En el run full, esta etapa operó a partir de un snapshot ya generado, correspondiente al corte 20260225_1928. Su función fue fijar una versión estable de los archivos de origen y evitar que el material cambiara a mitad del proceso. Metodológicamente, congelar los insumos es importante porque define un punto de referencia común: cualquier resultado posterior puede ser remitido al mismo universo de archivos de partida. Esta etapa permitió además inventariar los archivos disponibles y asociarlos con los módulos esperados, transformando un conjunto disperso de planillas en un corpus delimitado y auditable.

Una vez estabilizado el punto de partida, el siguiente problema fue la heterogeneidad estructural entre módulos. Para resolverlo se construyeron mappings específicos por módulo. Estos mappings definieron cómo traducir columnas raw a una estructura canónica común. Su función fue decisiva: sin ellos, integrar la información de módulos distintos habría implicado suponer equivalencias de manera implícita. El mapping, en cambio, obligó a declarar de forma explícita qué columna se consideraba equivalente a qué campo canónico, qué columnas eran indicadores, cuáles eran notas y cómo se trataban cabeceras especiales, vacías o duplicadas. La validación de mappings mostró coincidencia completa entre referencias esperadas y cabeceras resueltas para los once módulos, sin columnas huérfanas al cierre de esta etapa.

Definido el puente entre la estructura de origen y la estructura esperada, se construyó la tabla core por módulo. La tabla core puede entenderse como la capa base de cada módulo: reúne los campos esenciales que permiten identificar y describir cada registro en una forma homogénea. Su creación respondió a una necesidad analítica básica: antes de separar indicadores, notas o componentes específicos, había que asegurar que todas las observaciones compartieran un es-

queleto común. Esta fase permitió detectar tempranamente vacíos en campos relevantes y grupos duplicados de row_id, lo que sirvió como diagnóstico para las correcciones posteriores. En el resumen QC previo al patch, por ejemplo, se observaban duplicados de row_id en varios módulos, incluyendo Salud, HSE, DDHH y EDS, y presencia de cita vacía en módulos como PAZ y HSE.

El siguiente problema apareció en relación con artefactos visuales heredados de los archivos de origen, especialmente celdas combinadas y filas sin contenido analítico. Cuando una planilla usa celdas combinadas, es habitual que la información aparezca explícitamente sólo en la primera fila del bloque y quede vacía en las siguientes. Para lectura humana eso no es un problema, porque el ojo interpreta la continuidad. Para el procesamiento automatizado, en cambio, esos vacíos pueden desestabilizar la normalización de campos como documento y, por extensión, la construcción de llaves. Por ello se aplicó un patch de fill-down restringido por país y ordenado por fila de origen. En paralelo, se excluyeron filas con cita vacía cuando éstas actuaban como separadores o ruido y no como observaciones analíticas. Al mismo tiempo, se creó row _uid como llave operativa interna, capaz de preservar la trayectoria exacta de cada fila incluso cuando row_id todavía no era único. Después de esta etapa, los vacíos en documento quedaron resueltos y los vacíos en cita fueron eliminados de la base operativa.

Una vez estabilizada la capa core, el pipeline separó tres componentes: variables núcleo, indicadores y notas. Esta división respondió a una necesidad metodológica clara. No toda la información cumple la misma función analítica, y mezclarla en una sola tabla complica tanto el control como la interpretación. Separar core, indicators y notes permitió, por un lado, recodificar y auditar indicadores sin alterar variables estructurales y, por otro, conservar el material textual en una base paralela pero alineada. La verificación central de esta fase fue que los tres componentes mantuvieran correspondencia perfecta por row_uid dentro de cada módulo.

La etapa de recodificación de indicadores resolvió un problema distinto: la falta de un dominio comparativo común. Si los módulos usan marcas distintas para expresar presencia, ausencia, etiquetas o codificaciones parciales, no es posible sumar, comparar o auditar indicadores entre sí. Por ello se recodificaron al dominio {0,1,NA}, aplicando reglas globales y, cuando fue necesario, overrides por columna para variables cuya lógica de codificación estaba basada en etiquetas no vacías más que en marcas binarias convencionales. Gracias a esta etapa, los indicadores quedaron listos para uso comparativo y el pipeline pudo auditar explícitamente la

ausencia de valores inválidos fuera del dominio esperado. Del mismo modo, los residuos no binarios que no podían interpretarse de manera defendible como presencia o ausencia fueron preservados como NA.

La deduplicación trazable fue el paso siguiente. Aunque la corrección de vacíos y la separación por componentes habían ordenado la estructura, aún persistían grupos duplicados de row_id en algunos módulos. En vez de resolver esto de forma arbitraria, se adoptó una regla determinista: mantener como ganador el registro con menor source_row_number y filtrar los perdedores de manera consistente en core, indicators y notes mediante row_uid. Esta decisión fue clave porque permitió estabilizar el conteo sin sacrificar auditabilidad. En esta etapa, por ejemplo, Salud pasó de 969 a 909 registros, eliminando grupos duplicados que habrían afectado el conteo final.

2.4.1. Reducción de registros por módulo durante la consolidación

Un resultado importante del proceso de consolidación fue la reducción controlada de registros duplicados. Esta disminución no implicó pérdida arbitraria de información sustantiva, sino la eliminación trazable de observaciones que pertenecían a grupos duplicados de row_id y que, de mantenerse, habrían generado sobreconteo o ambigüedad analítica.

En términos operativos, cada módulo partió con un número inicial de registros (n_before) y terminó con un número final de registros válidos (n_after) luego de aplicar la regla de deduplicación determinista. La diferencia entre ambos valores corresponde a los registros eliminados en esta etapa. Los registros que quedaron fueron aquellos seleccionados como observación canónica dentro de cada grupo duplicado, siguiendo la regla de conservar el caso con menor source_row_number y filtrar consistentemente los restantes mediante row_uid.

2.4.2. Tabla 1. Registros antes y después de la deduplicación por módulo

Módulo	Registros antes (n_before)	Registros después (n_after)	Registros eliminados	Grupos duplicados antes	Grupos duplicados después
DDHH	670	658	12	12	0

Módulo	Registros antes (n_before)	Registros después (n_after)	Registros eliminados	Grupos duplicados antes	Grupos duplicados después
ECM	447	442	5	5	0
EDS	1409	1399	10	8	0
ESI	259	259	0	0	0
Género	316	316	0	0	0
HSE	1542	1526	16	16	0
HSE Respeto	441	439	2	1	0
PAZ	254	253	1	1	0
Pilares actitudinal	1006	989	17	17	0
Pilares cognitivo	221	220	1	1	0
Salud	969	909	60	50	0
Total	7534	7410	124	111	0

La tabla muestra que la reducción total fue de 124 registros, pasando de 7.534 observaciones previas a la deduplicación a 7.410 observaciones en la base final. Esta disminución se explica por la resolución de 111 grupos duplicados de row_id, todos eliminados al cierre de la etapa.

Con las tablas ya estabilizadas y deduplicadas, se construyó el master final. Esta etapa consistió en unir uno a uno las tablas estructurales e indicadores por row_uid y concatenar luego los once módulos en un único conjunto consolidado. El resultado de esta operación fue la base erce_m1_full_master.csv, junto con su contraparte erce_m1_full_notes.csv. Metodológicamente, este fue el momento en que la información dejó de estar organizada por módulos aislados y pasó a constituir un universo único, comparable y trazable. Los chequeos de cardinalidad confirmaron que en todos los módulos el join fue 1:1 y que no se generaron duplicados posteriores a la unión.

En esta base final, cada fila representa una evidencia o cita curricular específica, ubicada en un documento, módulo y página determinados. Sobre esa unidad de análisis se organizan vari-

ables núcleo, variables de contexto, variables de trazabilidad e indicadores recodificados. Esta precisión es importante porque permite interpretar correctamente la estructura del conjunto: la base no resume documentos completos ni módulos enteros, sino evidencias puntuales, comparables y rastreables.

El pipeline cerró con una fase de control de calidad final, descriptivos univariados, elaboración de codebook y chequeo de entrega. Esta etapa tuvo una doble función. Por una parte, validó que el resultado final cumpliera condiciones mínimas de integridad: ausencia de duplicados globales, control de faltantes esenciales, dominio válido de indicadores y consistencia entre master y notes. Por otra, produjo salidas interpretables que permiten leer cómo quedó distribuida la base por módulo, país, grado, asignatura y densidad de indicadores. Finalmente, el proceso se integró en un reporte final y una verificación de entrega, asegurando que no sólo existiera una base consolidada, sino también un conjunto de documentos y evidencias que permiten comprenderla, revisarla y reproducirla.

2.5. Problemas detectados y soluciones adoptadas

El primer gran tipo de problema fue la heterogeneidad de cabeceras. Algunos módulos presentaban nombres de columnas distintos para funciones equivalentes; otros incluían columnas vacías, duplicadas o con etiquetas repetidas. El riesgo de este tipo de heterogeneidad es sencillo de entender: si dos columnas representan lo mismo pero no se reconocen como equivalentes, la información queda fragmentada; si una sola etiqueta aparece duplicada con contenidos distintos, la integración puede mezclar elementos no equivalentes. La solución adoptada fue el mapping explícito por módulo, apoyado en resolución de cabeceras y perfiles de columnas. La alternativa descartada fue una integración implícita por coincidencia parcial de nombres o una revisión exclusivamente manual sin reglas formalizadas.

El segundo problema fue la presencia de celdas combinadas y vacíos estructurales. Aquí el riesgo no era la falta real de información, sino la apariencia de falta. Cuando el nombre de un documento se muestra una sola vez en una planilla y queda visualmente extendido a las filas siguientes, el procesamiento automatizado lee vacíos donde el lector humano ve continuidad. Si esos vacíos se mantienen, se desestabiliza la construcción de llaves y se fragmenta artificialmente la unidad documental. La decisión adoptada fue aplicar fill-down de documento dentro de cada p

ais, conservando el orden de origen. Se descartó tanto dejar los vacíos intactos como aplicar un relleno global sin fronteras, porque ambas opciones introducían riesgos analíticos mayores.

El tercer problema fue la existencia de filas separadoras o sin contenido analítico, identificables por cita vacía. El riesgo aquí consistía en contar como observación una fila que en realidad funcionaba como borde visual, subtítulo o espacio en blanco dentro de la planilla. La solución fue excluir esas filas del core analítico y exportarlas como evidencia de control de calidad. Se descartó mantenerlas y confiar en que el ruido se corregiría más adelante, porque eso habría contaminado etapas posteriores con registros que no correspondían a la unidad analítica definida.

El cuarto problema fueron los duplicados de row_id. Incluso después de estabilizar la estructura, algunos módulos mantenían grupos duplicados. Si esos duplicados no se resuelven, la base puede sobrecontar evidencias o generar joins ambiguos. La decisión adoptada fue una deduplicación determinista, con criterio de ganador explícito y trazabilidad mediante row_uid. Se descartó tanto conservar todos los duplicados como elegir un ganador de manera arbitraria por orden de archivo, porque ambas alternativas comprometían la reproducibilidad.

El quinto problema surgió en la recodificación de indicadores. Algunas columnas usaban marcas binarias convencionales; otras usaban etiquetas no vacías, cadenas tipo tupla o anotaciones breves que funcionaban de hecho como presencia de un rasgo. El riesgo en este punto era doble: perder señal válida si se trataban esas marcas como residuales o, a la inversa, forzar como presencia lo que era una anotación descriptiva ambigua. La solución fue combinar reglas globales con overrides por columna, documentando dónde un valor no vacío equivalía a presencia y dónde esa equivalencia no era defendible. Se descartó enumerar manualmente todas las variantes de tokens positivos o dejar como NA señales sistemáticas que sí tenían valor codificado consistente.

Un sexto problema apareció en el módulo EDS. Allí se habían detectado columnas unnamed con patrones de distribución compatibles con marcas analíticas y también residuos no binarios concentrados en una fila específica. La solución fue separar esas columnas entre candidatas a indicadores y notas cuando correspondía, y preservar como NA aquellos residuos que no podían leerse como marcas binarias defendibles. La alternativa descartada fue forzar esos residuos como presencia positiva.

Un séptimo problema, de naturaleza más específicamente documental, fue detectado en el módulo Salud. Durante la revisión final se identificó un subconjunto acotado de registros cor-

respondientes a Paraguay, específicamente al documento Paraguay 2 Programa Tercer_grado .pdf en la página 104, cuyos textos asociados a la cita presentaban problemas severos de legibilidad y codificación de caracteres. Estos registros sí presentaban activación de indicadores, pero no podían ser interpretados ni auditados sustantivamente de manera confiable en su estado actual. El riesgo de tratarlos como observaciones completamente estables era introducir en la base analítica una señal positiva cuya evidencia textual seguía siendo deficiente o dudosa. En esta etapa, sin embargo, estos casos no fueron eliminados, sino mantenidos temporalmente por razones de trazabilidad, a la espera de que la información pueda ser corregida o completada si se logra recuperar la referencia correspondiente.

Finalmente, hubo excepciones acotadas en variables de completitud, especialmente en pa gina, incluyendo el caso ya identificado en PAZ y otros casos observados en ESI y Salud. La fortaleza del proceso no radica en haber eliminado toda imperfección, sino en haber distinguido entre errores estructurales, ruido removible y excepciones legítimamente acotadas o pendientes de revisión.

2.6. Documentación de decisiones de curatoría y procesamiento de datos

Uno de los rasgos metodológicamente más importantes del proceso fue que las decisiones no quedaron implícitas en los scripts, sino explícitamente registradas en un decision_log. Esta documentación permitió transformar reglas operativas en decisiones defendibles. No se trató simplemente de “hacer que el pipeline funcionara”, sino de justificar por qué el modo en que funciona es razonable.

El decision_log muestra, además, que la consolidación no fue una secuencia improvisada de correcciones, sino un proceso acumulativo de aprendizaje metodológico. Las decisiones del piloto cumplieron una función exploratoria y de prueba de reglas; las decisiones del run full extendieron, formalizaron o generalizaron esos criterios al conjunto completo de once módulos. En este sentido, el registro de decisiones documenta no sólo qué se hizo, sino también cómo ciertos problemas fueron comprendidos progresivamente y luego estabilizados como reglas de procesamiento.

En términos analíticos, las decisiones registradas pueden distinguirse en tres grandes grupos. El primero corresponde a decisiones estructurales. Son aquellas que definen cómo representar la información: mappings por módulo, normalización de cabeceras, creación de tablas core y construcción de llaves operativas. Aquí se ubican, por ejemplo, TRIAL-EDS-001, TRIAL-ECM-001, TRIAL-DDHH-001 y el conjunto de decisiones FULL-MAP-* para ESI, Género, HSE, HSE Respeto, PAZ, Pilares actitudinal, Pilares cognitivo y Salud. Estas decisiones responden a la pregunta de cómo traducir el desorden original a una estructura analítica estable.

El segundo grupo corresponde a decisiones defensivas. Aquí se ubican reglas como el fill-down restringido, la exclusión de filas con cita vacía y la deduplicación determinista. Este grupo incluye TRIAL-CORE-002, TRIAL-CORE-003, TRIAL-DEDUPE-001, FULL-CORE-002, FULL-CORE-003 y FULL-DEDUPE-001. Se trata de decisiones orientadas a prevenir riesgos metodológicos concretos: sobreconteo, pérdida de trazabilidad, fragmentación de unidades o joins ambiguos. Su función no es reinterpretar el contenido, sino proteger la integridad del proceso.

El tercer grupo corresponde a decisiones de honestidad semántica. Estas aparecen cuando la evidencia disponible no permite forzar una clasificación binaria o inequívoca sin introducir una distorsión. Aquí destacan TRIAL-RECODE-001, TRIAL-RECODE-EDS-002, FULL-RECODE-MULTI-TAGCOLS-001 y FULL-RECODE-RESIDUAL-EDS-001. En estos casos, el criterio no fue maximizar artificialmente la completitud, sino preservar la relación defendible entre codificación y significado. El caso residual de EDS es el ejemplo más claro, pero también debe incluirse aquí el tratamiento de los casos documentales en revisión en Salud: aunque esos registros presentaban indicadores, la evidencia textual asociada no era legible de manera defendible en su estado actual, por lo que se optó por mantenerlos trazados y explícitamente observados, en lugar de cerrar prematuramente su tratamiento con una decisión no verificable.

El valor de esta documentación radica en que vuelve el proceso revisable. Una contraparte puede no sólo observar el resultado final, sino reconstruir el razonamiento que llevó a él. En ese sentido, la trazabilidad no se limita a las filas de la base, sino que alcanza también las decisiones que la hicieron posible.

2.6.1. Tabla 2. Síntesis de decisiones metodológicas

Tipo de decisión	Ejemplos del decision_log	Función metodológica
Estructurales	TRIAL-EDS-001, TRIAL-ECM-001, TRIAL-DDHH-001, FULL-MAP-ESI-001, FULL-MAP-GENERO-001, FULL-MAP-HSE-001, FULL-MAP-HSE_RESPETO-001, FULL-MAP-PAZ-001, FULL-MAP-PILARES_ACTITUDINAL-001, FULL-MAP-PILARES_COGNITIVO-001, FULL-MAP-SALUD-001	Definir mappings, resolver cabeceras, separar indicadores y notas, y construir la estructura canónica del corpus
Defensivas	TRIAL-CORE-002, TRIAL-CORE-003, TRIAL-DEDUPE-001, FULL-CORE-002, FULL-CORE-003, FULL-DEDUPE-001	Evitar ruido, sobreconteo, pérdida de trazabilidad y ambigüedad en joins mediante fill-down controlado, exclusión de separadores y deduplicación determinista
Honestidad semántica	TRIAL-RECODE-001, TRIAL-RECODE-EDS-002, FULL-RECODE-MULTI-TAGCOLS-001, FULL-RECODE-RESIDUAL-EDS-001	Preservar una relación defendible entre señal codificada y significado, evitando forzar recodificaciones no binarias o ambiguas

3 Libro de códigos y descriptivos univariados

3.1. Libro de códigos y estructura de la base

Lectura general de la base

La base consolidada final quedó organizada en torno a dos productos principales: una base master y una tabla notes. Ambas comparten el mismo universo de observaciones y el mismo número total de filas, pero cumplen funciones distintas y complementarias dentro del proceso de análisis.

La unidad de análisis de la base es explícita: cada fila corresponde a una evidencia o cita curricular específica, ubicada en un documento, módulo y página determinados. Esto significa que la base no resume documentos completos ni países en su conjunto, sino evidencias puntuales organizadas bajo una estructura común y comparable.

Estructura general de productos

Producto	Función principal	Contenido	Uso analítico
master	Base analítica principal	Variables núcleo, contexto, trazabilidad e indicadores ind_*	Análisis cuantitativo
notes	Tabla auxiliar alineada	Notas, comentarios o texto complementario cuando existe	Revisión y respaldo

Producto	Función principal	Contenido	Uso analítico
codebook_full.csv	Libro de códigos resumido	Variables, grupos y descripciones básicas	Consulta técnica rápida
codebook_master_explicito.csv	Libro de códigos detallado	Definición, origen raw, reglas de construcción y calidad observada	Auditoría detallada

Tipos de variables

La base master concentra cuatro grandes grupos de variables, cada uno con una función analítica específica:

Grupo de variables	Función	Ejemplos
Núcleo	Identifican y ubican la observación	module, pais, documento, pagina, cita
Contexto	Describen características sustantivas del registro	grado_norm, materia_norm, campo_norm, categoria, nivel, subdimension, subcat
Trazabilidad	Permiten reconstruir el origen exacto de la fila	id_raw, row_id, row_uid, source_file, source_sheet, source_row_number
Indicadores	Registran la presencia o ausencia de criterios analíticos	variables ind_*

Interpretación de las principales variables

Variables núcleo

Las variables núcleo permiten ubicar e identificar cada evidencia curricular dentro del corpus consolidado. Entre ellas se encuentran module, pais, documento, documento_norm, pagina, pagina_norm, cita, cita_norm, grado_norm, materia_norm y campo_norm.

Estas variables no cumplen todas la misma función. Algunas forman parte del núcleo mínimo de identificación de la observación, mientras que otras agregan información contextual cuando la fuente original lo permite.

Variable	Función	Observación de lectura
module	Identifica el módulo temático	Está completa en toda la base
pais	Identifica el país del documento	Está completa en toda la base
documento	Nombre original del documento fuente	Está completo en toda la base
pagina	Página reportada en la base cruda	Tiene excepciones puntuales documentadas
cita	Texto original de la evidencia curricular	Está completa en toda la base
documento_norm / pagina_norm / cita_norm	Versiones normalizadas para joins y llaves	Se usan para comparación y deduplicación
grado_norm, materia_norm, campo_norm	Variables de clasificación contextual	Su cobertura depende del módulo

Completitud del núcleo mínimo de identificación

Desde el punto de vista metodológico, el núcleo mínimo de identificación de una observación está compuesto por module, pais, documento, pagina y cita. Estas variables permiten ubicar una evidencia concreta dentro del corpus y sostener su lectura analítica en condiciones de trazabilidad.

En la base consolidada final, module, pais, documento y cita presentan completitud total, mientras que pagina registra un nivel de faltantes muy bajo y acotado a excepciones ya documentadas. Esto confirma que la base quedó estructurada de manera suficientemente robusta para etapas posteriores del análisis.

Campo esencial	Lectura general
module	Sin faltantes
pais	Sin faltantes

Campo esencial	Lectura general
documento	Sin faltantes
cita	Sin faltantes
pagina	Faltantes puntuales, bajos y documentados

Llaves y trazabilidad

Las variables de trazabilidad permiten reconstruir el recorrido exacto de una fila desde el m aster final hasta su origen en el snapshot congelado.

Variable	Qué representa	Función metodológica
id_raw	Identificador original de la fila cruda	Conserva referencia original
row_id	Llave canónica construida con campos normalizados	Resume la identidad analítica del registro
row_uid	Llave operativa única (row_id::source_row_number)	Asegura unicidad operativa en el pipeline
source_snapshot_id	Snapshot congelado de origen	Permite reproducibilidad
source_file	Archivo Excel de origen	Permite volver al documento fuente
source_sheet	Hoja de origen	Localiza la evidencia dentro del archivo
source_row_number	Número de fila de origen	Permite auditoría fina

En particular, row_id resume la identidad canónica de la fila una vez normalizados sus campos centrales. row_uid, en cambio, representa su identidad operativa única dentro del pipeline y fue clave para las etapas de patch, split, recode, dedupe y join.

Indicadores ind_*

Los indicadores constituyen la capa analítica propiamente tal. Fueron recodificados a un dominio común {1,0,NA} para asegurar comparabilidad entre módulos.

Valor	Significado
1	Presencia o marca positiva del criterio
0	Negativo explícito en la fuente raw
NA	Vacío, no observado, no aplicable o señal no interpretable

Los indicadores no ingresan directamente al master desde la base cruda. Primero se extraen a tablas intermedias por módulo, luego se recodifican, posteriormente pasan por deduplicación y recién entonces se incorporan al master final.

Nota sobre la tabla notes

La tabla notes funciona como una estructura auxiliar alineada con la base master. Su objetivo es conservar, cuando existen, notas, comentarios o campos textuales complementarios asociados a una observación, sin romper su correspondencia con la fila analítica principal.

Que master y notes tengan el mismo número de filas no implica que ambas contengan el mismo volumen de contenido sustantivo. Lo que esta equivalencia asegura es **alineación estructural**: cuando existe información textual complementaria, ésta puede vincularse sin desajustes con la observación correspondiente.

3.2. Descriptivos univariados de la base consolidada

Panorama general

La producción de descriptivos univariados al cierre de M1 cumple una función doble. Por una parte, permite caracterizar el estado de la base resultante. Por otra, constituye un control adicional sobre la plausibilidad del resultado final.

Indicador global	Valor
Filas en master	7410
Filas en notes	7410
Duplicados globales de row_id	0
Duplicados globales de row_uid	0
Valores inválidos en indicadores	0

Indicador global	Valor

Estas cifras muestran que la base consolidada alcanzó condiciones mínimas de integridad para ser utilizada como soporte del análisis posterior. En términos operativos, al cierre de M1 se cuenta con una base alineada, sin duplicados globales de llaves y con indicadores contenidos en un dominio válido y comparable.

Cobertura por módulo

La base final reúne 7.410 observaciones distribuidas entre once módulos. La distribución es desigual, pero ello no constituye un error: refleja la heterogeneidad original del corpus y el hecho de que los módulos no tenían el mismo volumen de evidencia en las fuentes de origen.

Módulo	N filas
HSE	1526
EDS	1399
Pilares actitudinal	989
Salud	909
DDHH	658
ECM	442
HSE Respeto	439
Género	316
ESI	259
PAZ	253
Pilares cognitivo	220

Los mayores volúmenes se concentran en HSE, EDS y Pilares actitudinal, mientras que Pilares cognitivo, PAZ y ESI presentan tamaños comparativamente menores. En esta etapa, el interés principal de esta distribución no es aún sustantivo, sino metodológico: confirma que la consolidación preservó la estructura del corpus sin forzar una uniformidad artificial entre módulos.

Cobertura por país

Los descriptivos por país permiten verificar que la base final ya puede ser organizada territorialmente de manera consistente. Esto es relevante porque uno de los usos esperados de la base en las siguientes fases del proyecto es la comparación entre países.

En esta entrega, la lectura de la distribución por país no busca todavía producir conclusiones sustantivas, sino mostrar que la información quedó disponible en una forma que admite desagregación nacional sin perder coherencia estructural. La tabla siguiente resume la cobertura total consolidada del corte, agregando los módulos en una sola vista para facilitar la lectura del conjunto.

También conviene notar que en la salida observada persisten algunas grafías diferenciadas para un mismo país, como Perú y Peru, o México y Mexico. En esta instancia se mantienen separadas para preservar trazabilidad respecto de la exportación efectivamente generada en este corte.

País	N filas
Guatemala	975
Nicaragua	929
Venezuela	640
República Dominicana	624
Ecuador	493
Costa Rica	448
Colombia	390
Honduras	341
Chile	330
Brasil	297
El Salvador	285
Paraguay	285
Perú	269
Bolivia	243
Panamá	236
México	195
Argentina	188
Cuba	120

País	N filas
Uruguay	77
Mexico	23
Peru	22

Cobertura por grado

La distribución por grado permite observar hasta qué punto la base consolidada puede organizarse por nivel educativo cuando esa información se encuentra presente en los módulos de origen y fue normalizada durante el procesamiento.

La presencia de faltantes en grado_norm no debe interpretarse automáticamente como un problema de calidad, ya que en muchos casos refleja que la variable no estaba disponible, no era homogénea entre módulos o no era pertinente para ciertos tipos de evidencia curricular. Por ello, esta tabla debe leerse como una descripción de cobertura efectiva y no como un criterio único de calidad del dataset.

Para facilitar la lectura, se presenta a continuación una síntesis consolidada de la distribución observada en grado_norm. Se conservan las etiquetas tal como aparecen en la salida del corte, incluyendo formas numéricas y textuales; únicamente el valor vacío se muestra como (vacío).

grado_norm	N filas
(vacío)	4995
6	803
3	695
sexto_grado	345
tercer_grado	276
unico_documento	157
sexto_ano	62
tercer_ano	32
quinto_y_sexto_grado	19
tercer_y_cuarto_grado	16
4	4

grado_norm	N filas
tercero	4
sexto	2

En conjunto, esta distribución confirma que la base final permite desagregar una parte importante del corpus por nivel educativo, aunque con intensidades distintas según módulo. Esto no debilita la calidad de la base; más bien muestra que la disponibilidad de esta dimensión depende de la estructura original de los insumos y del tipo de evidencia curricular codificada en cada caso.

Cobertura por materia

La organización por materia o asignatura muestra que la base final puede ser utilizada para lecturas curriculares más finas, desagregadas por disciplina. Esto fortalece su utilidad analítica para las siguientes etapas del proyecto.

Como ocurre con otras variables contextuales, la cobertura de materia_norm depende de la estructura de los módulos y de las condiciones en que esa variable estaba presente en los insumos originales.

Para evitar una tabla excesivamente extensa, se presenta una síntesis consolidada de las categorías con mayor frecuencia en materia_norm. Se conserva la lógica del corte observado: el valor vacío se muestra como (vacío) para facilitar la lectura, mientras que etiquetas como n_a se mantienen tal como aparecen en la salida exportada.

materia_norm	N filas
(vacío)	3422
ciencias_naturales	415
n_a	383
ed_fisica	356
ciencias_sociales_formacion_etica_y_ciudadana	304
ciencias_sociales	299
ciencias	294
educacion_fisica	159

materia_norm	N filas
derechos_y_dignidad_de_las_mujeres	140
matematica	138
cs_social	129
ciencias_naturales_y_tecnologia	124
lengua	93
estudios_sociales_y_educacion_civica	90
ciencias_de_la_naturaleza_y_tecnologia	89
ciencias_de_la_naturaleza	75
creciendo_en_valores	65
gral	54
estudios_sociales	51
espanol	49

En conjunto, esta distribución muestra que la base ya puede organizarse por asignatura o campo disciplinar, aunque con nomenclaturas heterogéneas heredadas de los insumos originales. Esto no debilita la utilidad analítica del conjunto; al contrario, hace visible una dimensión de variación que podrá ser afinada o agrupada en etapas posteriores si el análisis comparativo así lo requiere.

Densidad de indicadores por fila

La densidad de indicadores por fila muestra diferencias entre módulos. Esto expresa que no todos comparten la misma arquitectura de codificación ni el mismo nivel de granularidad analítica.

Módulo	Promedio de activaciones por fila
EDS	3.14
HSE	2.47
Salud	2.43
Género	2.35
HSE Respeto	2.25
PAZ	1.80

Módulo	Promedio de activaciones por fila
DDHH	1.50
ESI	1.44
ECM	1.26
Pilares actitudinal	1.20
Pilares cognitivo	0.47

EDS presenta la mayor densidad promedio de activaciones por fila, seguido por HSE y Salud. En el extremo inferior, Pilares cognitivo muestra la menor densidad. Esta variación no debe leerse como inconsistencia, sino como resultado de diferencias reales en la estructura de codificación de cada módulo. Precisamente por ello fue necesario recodificar los indicadores a un dominio común antes de integrarlos en una única base.

Lectura de faltantes

La lectura de los descriptivos debe distinguir entre variables esenciales y variables contextuales.

Tipo de variable	Comportamiento esperado
Variables esenciales (module, pais, documento, pagina, cita)	Deben tener completitud muy alta
Variables contextuales (grado_norm, materia_norm, campo_norm, categoria, nivel, subdimension, subcat)	Pueden presentar faltantes altos sin implicar un error

En otras palabras, no todos los faltantes deben interpretarse como problemas de calidad. En muchos casos reflejan que la variable no existe en todos los módulos o que sólo es pertinente para ciertos tipos de evidencia curricular. Por ello, el juicio de calidad debe centrarse primero en el núcleo mínimo de identificación de las observaciones y no en exigir cobertura homogénea a variables contextuales que dependen de la estructura de cada módulo.

3.3. Anexo. Codebook explícito del master final

Fuente principal: runs/erce_m1_full/outputs/master/erce_m1_full_master.csv

Export complementario: runs/erce_m1_full/outputs/reports/codebook_master_explicito.csv

A.1. Cómo leer este anexo

En este anexo se documenta, para cada variable, su grupo funcional, definición, dominio esperado, módulos donde aparece, origen raw, regla de construcción y calidad observada en este corte.

A.2. Variables núcleo y de trazabilidad

variable	variable_group	descripcion	dominio	calidad_en_este_corte
module	nucleo	Módulo temático al que pertenece la fila en el master final	Texto libre normalizado o valor técnico derivado	pct_missing=0.0; n_unique=11; ejemplos=hse | eds | pilares_actitudinal
pais	nucleo	País del documento curricular de origen	Texto libre normalizado o valor técnico derivado	pct_missing=0.0; n_unique=21; ejemplos=Guatemala | Nicaragua | Venezuela
documento	nucleo	Nombre original del documento curricular en la fuente cruda	Texto libre normalizado o valor técnico derivado	pct_missing=0.0; n_unique=293; ejemplos=Adecuacion-Curricular-Primaria-_2023.pdf | Guatemala CNB_3er_grado.pdf

variable	variable_group	descripcion	dominio	calidad_en_este_corte
pagina	nucleo	Página reportada en la base cruda	Texto libre normalizado o valor técnico derivado	pct_missing=0.0135; n_unique=408; ejemplos=25 | 28 | 26
cita	nucleo	Texto raw de la evidencia curricular leída desde la base cruda	Texto libre normalizado o valor técnico derivado	pct_missing=0.0; n_unique=6023
row_id	trazabilidad	Llave canónica hash de la fila, construida con campos normalizados	Texto libre normalizado o valor técnico derivado	pct_missing=0.0; n_unique=7410
source_file	trazabilidad	Archivo Excel de origen en la capa raw	Texto libre normalizado o valor técnico derivado	pct_missing=0.0; n_unique=9
source_sheet	trazabilidad	Hoja Excel de origen en la capa raw	Texto libre normalizado o valor técnico derivado	pct_missing=0.0; n_unique=5
source_row_number	trazabilidad	Número de fila del lector en la hoja de origen	Texto libre normalizado o valor técnico derivado	pct_missing=0.0; n_unique=1543
row_uid	trazabilidad	Llave operativa única para joins internos	Texto libre normalizado o valor técnico derivado	pct_missing=0.0; n_unique=7410

A.3. Indicadores

variable	variable_group	descripcion	dominio	calidad_en_este_corte
ind_ddhh	indicador	Marca presencia temática de derechos humanos	{1, 0, NA}	pct_missing=95.5466; n_unique=2
ind_democracia	indicador	Marca presencia temática de democracia	{1, 0, NA}	pct_missing=96.0999; n_unique=2
ind_tag_ecm	indicador	Tag temático ECM (Educación para la Ciudadanía Mundial)	{1, 0, NA}	pct_missing=93.1444; n_unique=2
ind_cc	indicador	Marca presencia de cambio climático	{1, 0, NA}	pct_missing=99.3522; n_unique=2
ind_salud_mental	indicador	Marca presencia de salud mental	{1, 0, NA}	pct_missing=99.9595; n_unique=2

A.4. Origen raw y reglas de recodificación de indicadores

Como complemento del resumen anterior, se generó un export específico que documenta, para cada indicador y módulo, la columna raw de origen, la regla de mapping y el override de recodificación aplicado cuando corresponde. Este insumo permite auditar con mayor detalle cómo se tradujo cada marca observada en la fuente cruda al dominio analítico común {1,0,NA}.

La cobertura de este export no es uniforme por módulo, porque depende de la cantidad y diversidad de indicadores efectivamente mapeados en cada caso. En el corte utilizado para este informe, la distribución resumida es la siguiente:

Módulo	N indicadores documentados
salud	21
hse_respeto	20
hse	19
eds	15
genero	14
paz	12
ddhh	8
ecm	7
pilares_actitudinal	4
esi	3
pilares_cognitivo	2

Esta información es metodológicamente relevante por dos razones. En primer lugar, permite verificar que los indicadores no fueron incorporados de manera opaca al master final, sino a partir de columnas raw identificables y reglas de recodificación explícitas. En segundo lugar, permite distinguir entre los casos en que se aplicó una recodificación binaria convencional (x/si/1 => 1, no/0 => 0) y aquellos en que fue necesario utilizar overrides como non_empty_is_positive, especialmente en variables tipo tag o en columnas cuya presencia no estaba expresada como binario puro.

Para no sobrecargar el cuerpo del codebook, la tabla siguiente presenta una selección representativa de indicadores, módulos y reglas. El detalle exhaustivo se conserva en runs/erce_m1_full/outputs/reports/codebook_indicadores_origen_por_modulo.csv.

Variable	Módulo	Columna raw de origen	Regla / override	Descripción
ind_a2030_alinea	ecm	A2030-se alinea	si/x=>1; no=>0; else NA	Marca alineación explícita con Agenda 2030.

Variable	Módulo	Columna raw de origen	Regla / override	Descripción
ind_a2030_alinea_guess	eds	blank_col_pos_36	si/x=>1; no=>0; else NA	Inferencia conservadora de alineación a Agenda 2030 desde columnas unnamed en EDS.
ind_adicciones	salud	adicciones	Override non_empty_is_positive	Marca presencia de adicciones.
ind_agencia	hse	Agencia	Regla general {x/si/1/true}=1, {no/0/false}=0, vacío=NA	Marca presencia de agencia.
ind_agencia	hse_respeto	Agencia	Regla general {x/si/1/true}=1, {no/0/false}=0, vacío=NA	Marca presencia de agencia.
ind_consumo_responsable	pilares_actitudinal	Consumo responsable	Regla general {x/si/1/true}=1, {no/0/false}=0, vacío=NA	Marca presencia de consumo responsable.
ind_ddhh	ddhh	ddhh	x/si/1=>1; no/0=>0; else NA	Marca presencia temática de derechos humanos.
ind_empodera_4	genero	4 empodera	Regla general {x/si/1/true}=1, {no/0/false}=0, vacío=NA	Marca presencia de empoderamiento.

Variable	Módulo	Columna raw de origen	Regla / override	Descripción
ind_paz_tematica	paz	Paz	Override non_empty_is_positive	Marca presencia de temática de paz.
ind_seguridad_personal	esi	Seguridad personal	Regla general {x/si/1/true}=1, {no/0/false}=0, vacío=NA	Marca presencia de seguridad personal.
ind_toma_decisiones_responsable	pilares_cognitivo	Toma de decisiones responsable	Regla general {x/si/1/true}=1, {no/0/false}=0, vacío=NA	Marca presencia de toma de decisiones responsable.
ind_tag_ecm	ecm	ECM	Override non_empty_is_positive	Tag temático ECM (Educación para la Ciudadanía Mundial).
ind_tag_ddhh	salud	DDHH	Override non_empty_is_positive	Tag temático DDHH.
ind_salud_mental	salud	salud mental	Override non_empty_is_positive	Marca presencia de salud mental.

La lectura combinada de este export y del codebook explícito permite entender no sólo qué significa cada indicador en el resultado final, sino también de dónde provino y bajo qué lógica fue recodificado. Esto fortalece la auditabilidad del proceso y hace posible reconstruir, con suficiente detalle, la relación entre la base cruda, las reglas intermedias y la base maestra consolidada.

Nota: El detalle exhaustivo de todas las variables está en runs/erce_m1_full/outputs /reports/codebook_master_explicito.csv, mientras que el detalle completo de origen

raw y recodificación de indicadores se encuentra en runs/erce_m1_full/outputs/reports/codebook_indicadores_origen_por_modulo.csv.