

KNO-019 · Datos, tablas y registros

KNO · Formas de conocer

Juan Seguro

v0.7 · ES

Todos los derechos reservados

Resumen del artículo

El trabajo científico deja tablas, hojas de cálculo, conjuntos de datos y archivos, pero la supervivencia de un fichero no indica por sí sola qué podrán aprender de él investigadores posteriores. ¿Qué debe conservar un registro para que, años después, alguien pueda interpretar responsablemente una observación, distinguir ausencia de no detección, identificar la versión analizada y reconstruir las transformaciones entre adquisición y publicación? Este artículo examina esas preguntas a través de estructura tabular, prácticas de hoja de cálculo, metadatos, procedencia, versionado, identificadores persistentes y la categoría ambigua de «datos brutos». Pregunta dónde termina el simple almacenamiento y comienza un registro científico auditable, cuánto contexto debe viajar con los valores numéricos y qué partes de la historia de producción de los datos deben seguir visibles para reanálisis, corrección y reutilización a largo plazo.

En 60 segundos

Los datos se convierten en evidencia duradera solo cuando su estructura y su historia siguen siendo interpretables. Las tablas facilitan algunas comparaciones, pero la comodidad de las hojas de cálculo también puede ocultar unidades, tipos, fórmulas o convenciones de valores ausentes. Los metadatos conservan el contexto necesario para entender qué significan los valores. La procedencia registra cómo se produjeron o transformaron los datos. El versionado importa porque citar «el conjunto de datos» es ambiguo si el objeto cambia, mientras que los identificadores persistentes estabilizan la referencia sin sustituir la documentación. Los valores ausentes deben representarse explícitamente porque ausencia, no detección y no aplicabilidad no son el mismo estado. Los «datos brutos» tampoco son necesariamente intactos: instrumentos y sistemas de adquisición pueden haber aplicado calibración o procesamiento. Preservar tiene éxito solo si futuros usuarios pueden reconstruir suficientemente la cadena desde observación hasta análisis. Un buen registro no es solo un archivo que sobrevive, sino uno cuyo significado, linaje y versión siguen siendo auditables.

Por qué importa esta pregunta

La ciencia moderna depende de registros que sobreviven al momento de la observación. Las mediciones se copian en tablas, se transforman con software, se combinan con metadatos, se depositan en repositorios y se citan en trabajos posteriores. Cada paso puede conservar información o destruir contexto silenciosamente. Un conjunto de datos puede estar numéricamente intacto y resultar inutilizable si faltan definiciones de variables, unidades, exclusiones o historia de procesamiento. El problema importa para reproducibilidad, reanálisis, corrección de errores y preservación a largo plazo. Entender los registros como objetos estructurados, versionados y con procedencia permite distinguir «los datos siguen existiendo» de la afirmación más fuerte de que otra persona todavía puede comprender qué significan y cómo llegaron a ser evidencia.

1. Un registro no es el acontecimiento

Un registro es una representación que deja un acontecimiento, una observación, una medición o una decisión. Puede conservar números, texto, imágenes, marcas temporales o categorías, pero no es idéntico a aquello que los produjo. **C01. Los registros amplían la memoria al hacer persistentes las observaciones, pero la persistencia no garantiza la interpretabilidad.**

Síntesis declarada: [3] y [5] respaldan la procedencia, los metadatos y la reconstrucción; no establecen directamente la cláusula sobre interpretabilidad.

Esta distinción importa porque los usuarios posteriores suelen encontrarse con el registro cuando el contexto original ya ha desaparecido. La tarea epistémica no consiste solo en almacenar contenido, sino en conservar suficiente información sobre su producción para que otra persona pueda comprender qué significa.

En qué punto estamos

Un registro representa un acontecimiento u observación; no es el acontecimiento mismo, de modo que preservar siempre depende de qué se codificó y qué quedó fuera. La afirmación es conceptual y no niega que los registros puedan ser evidencia fiable. La siguiente pregunta es estructural: ¿qué organización necesitan los datos para inspeccionar de forma consistente las relaciones entre observaciones?

2. Los datos necesitan estructura

Una colección de valores resulta más fácil de inspeccionar y analizar cuando su estructura es explícita. La formulación de datos ordenados de Wickham es una convención influyente: las variables ocupan columnas, las observaciones ocupan filas y los distintos tipos de unidades observacionales ocupan tablas distintas. [1]

C02. Una estructura tabular coherente puede reducir la ambigüedad y facilitar la transformación, el modelado y la visualización, pero es una convención para organizar datos, no una ontología universal del mundo. [1]

Una fila no es de manera natural una «observación» hasta que alguien define la unidad observacional. El mismo acontecimiento puede representarse al nivel de una persona, una visita, una lectura instrumental, una imagen, un ensayo o una transacción. La tabla incorpora, por tanto, una decisión de modelado.

En qué punto estamos

Los datos estructurados reducen ambigüedad al hacer explícitas variables, observaciones y valores, aunque ningún diseño tabular sea neutral respecto de operaciones posteriores. La ganancia epistémica es consistencia e inspeccionabilidad, no corrección automática. Una vez explícita la estructura, la siguiente pregunta es qué revelan bien las tablas y qué pueden ocultar al comprimir relaciones en filas y columnas.

3. Las tablas muestran unas relaciones y ocultan otras

Las tablas son poderosas porque alinean valores comparables. Permiten ordenar, filtrar, unir, agrupar y calcular. Sin embargo, una tabla puede ocultar la historia que produjo sus celdas. Una columna final puede contener un valor generado mediante varias transformaciones cuyos pasos intermedios no están presentes.

C03. La organización tabular facilita la comparación y el cálculo, pero la tabla por sí sola no tiene por qué conservar la procedencia de cada valor. Síntesis declarada: [2] respalda la organización tabular disciplinada y [3] respalda la procedencia; ninguna de las dos fuentes afirma directamente la cláusula de no conservación.

Por eso, una tabla limpia y un registro auditable no son lo mismo. Una buena organización ayuda al análisis; la procedencia ayuda a reconstruirlo.

En qué punto estamos

Las tablas exponen comparaciones, uniones y patrones con eficiencia, pero suprimen otras relaciones que pueden requerir representaciones distintas. Su utilidad depende de ajustar la estructura a la pregunta. Ese límite conduce a un peligro práctico: cuando los datos tabulares se manejan en hojas de cálculo flexibles, ¿qué comodidades pueden debilitar silenciosamente la integridad y auditabilidad del registro?

4. La comodidad de las hojas de cálculo puede crear fragilidad epistémica

Las hojas de cálculo son útiles porque combinan entrada de datos, cálculo y presentación. Esa misma flexibilidad también puede mezclar datos con formato, comentarios, varias variables en una sola celda, cálculos incrustados o códigos inconsistentes para valores ausentes. Broman y Woo recomiendan una organización rectangular simple, una sola pieza de información por celda, convenciones coherentes para valores ausentes y separación entre datos originales y cálculos. [2]

C04. Las prácticas de hoja de cálculo que mezclan observaciones originales, formato y cálculos derivados pueden hacer más propensas a error la interpretación y el análisis posteriores. [2]

No es un argumento contra las hojas de cálculo. Es un argumento a favor de tratar la estructura del registro como parte del método.

En qué punto estamos

Las hojas de cálculo son entornos de trabajo potentes, pero tipos mezclados, fórmulas ocultas, formato manual y celdas sobrecargadas pueden volver los datos frágiles o ambiguos. El riesgo no es que sean inválidas por naturaleza, sino que la comodidad oculte estructura. La siguiente pregunta es qué información debe acompañar a los valores para que futuros usuarios puedan interpretarlos correctamente.

5. Los metadatos conservan el significado que rodea a los valores

Los metadatos describen los datos. Pueden identificar variables, unidades, instrumentos, fechas, lugares, creadores, métodos, software, condiciones de acceso o vínculos con registros relacionados. El Research Data Framework de NIST trata los metadatos como esenciales para el uso, reutilización y preservación eficaces a lo largo del tiempo. [5]

C05. Los valores se vuelven más reutilizables cuando los metadatos conservan las definiciones y condiciones necesarias para interpretarlos. [5]

Una columna etiquetada `temperature` es incompleta si la escala, el sensor, el lugar, el momento y el procesamiento son relevantes para la pregunta. Los metadatos no eliminan mágicamente la incertidumbre; conservan el contexto necesario para razonar sobre ella.

En qué punto estamos

Los metadatos conservan el contexto alrededor de los valores: definiciones de variables, unidades, métodos, identificadores y demás información necesaria para interpretar el registro. Unos metadatos completos no garantizan ciencia válida, pero la falta de contexto puede inutilizar observaciones válidas. La siguiente pregunta es histórica: ¿cómo conservar no solo qué son los valores, sino cómo llegaron a existir?

6. La procedencia registra cómo llegaron a existir los datos

La familia W3C PROV define la procedencia en términos de entidades, actividades y agentes implicados en la producción de un dato u otro objeto. Está concebida para permitir descripciones interoperables de los orígenes y las transformaciones. [3]

C06. Los registros de procedencia pueden documentar de dónde proceden los datos, quién o qué actuó sobre ellos y qué procesos contribuyeron al objeto resultante. [3]

La procedencia es, por tanto, distinta de una etiqueta descriptiva. Un conjunto de datos puede tener nombres de variables excelentes y, aun así, no mostrar qué archivo fuente, ejecución instrumental, script o transformación produjo la versión analizada.

En qué punto estamos

La procedencia registra el linaje de los datos a través de adquisición, transformación y uso, permitiendo inspeccionar cómo se produjo un objeto. Mejora la reconstructibilidad sin demostrar que cada transformación haya sido apropiada. Si el linaje importa, aparece otro problema: cuando un conjunto de datos cambia con el tiempo, ¿qué estado exacto de ese objeto evolutivo identifica una cita posterior?

7. El versionado cambia aquello a lo que apunta una cita

Los datos pueden cambiar después de su creación inicial: se corrigen errores, se añaden registros, se revisan clasificaciones y se actualizan los procesos de tratamiento. Si una afirmación depende de un estado concreto de un conjunto de datos, citar solo el nombre general del conjunto puede ser insuficiente.

La Joint Declaration of Data Citation Principles subraya la especificidad y verificabilidad, incluida información suficiente para identificar una versión, un intervalo temporal o una porción granular concretos de los datos. [4]

C07. Cuando los datos cambian con el tiempo, la información de versión puede ser necesaria para determinar qué datos exactos sustentaron una afirmación. [4][5]

La afirmación de conocimiento queda vinculada no solo a una familia de datos, sino a una manifestación identificable de ella.

En qué punto estamos

El versionado distingue estados sucesivos de un objeto de datos, permitiendo que una referencia señale lo realmente analizado y no un objetivo móvil inespecífico. Una etiqueta de versión solo ayuda si esa versión se conserva e identifica. La siguiente pregunta es cómo contribuyen los identificadores persistentes y qué problemas de documentación no pueden resolver por sí solos.

8. Los identificadores persistentes resuelven un problema, no todos

Los identificadores persistentes ayudan a mantener estables las referencias cuando los objetos cambian de ubicación. Los principios de citación de datos recomiendan una identificación globalmente única, persistente y procesable por máquinas. [4]

C08. Un identificador persistente mejora la estabilidad referencial, pero por sí solo no establece la calidad de los datos, la corrección de la interpretación ni la relevancia probatoria. [4]

Un DOI puede decirnos qué objeto se citó. No puede decirnos si las mediciones eran válidas o si el análisis respondía a la pregunta pretendida. Identificación y evaluación son tareas epistémicas distintas.

En qué punto estamos

Los identificadores persistentes hacen las referencias más estables y resolubles por máquinas, pero no aportan por sí mismos metadatos, procedencia ni semántica de versiones ausentes. Resuelven identificación, no todo el problema de interpretabilidad. La siguiente pregunta vuelve al interior de la tabla: cuando una celda carece de valor ordinario, ¿cómo distinguir ausencia, cero, no detección o no aplicabilidad?

9. Los valores ausentes también son registros

La ausencia en una tabla puede significar muchas cosas: no medido, no aplicable, por debajo del límite de detección, perdido, rechazado, imposible, corrupto o realmente cero. Tratar estos estados como intercambiables puede cambiar el análisis.

C09. La ausencia de datos requiere representación explícita porque un campo vacío, un cero y un valor no observado pueden tener significados distintos. [2]

El esquema de codificación exacto depende del dominio. El punto metodológico general es que un registro debe distinguir estados diferentes cuando esas diferencias importan para la interpretación.

En qué punto estamos

Los valores ausentes son estados que requieren representación explícita, porque razones distintas para la ausencia pueden implicar análisis e interpretaciones diferentes. Codificar bien la ausencia no recupera información nunca observada, pero evita colapsar varios estados en uno. La siguiente pregunta es si la etiqueta habitual «datos brutos» identifica de manera fiable un punto de partida realmente intacto.

10. Los datos «brutos» no siempre son datos intactos

La expresión «datos brutos» suele sugerir una capa prístina anterior a la interpretación. En la práctica, los instrumentos pueden aplicar ya calibración, filtrado, compresión, umbrales o conversiones propietarias antes de que el investigador reciba un archivo. El RDaF de NIST trata explícitamente los instrumentos, las operaciones de procesamiento, los parámetros, los flujos de trabajo y la procedencia como partes relevantes de la gestión de datos. [5]

C10. El registro digital más temprano disponible puede contener ya decisiones de procesamiento, por lo que «bruto» no debe suponerse equivalente a «sin transformar».

Síntesis declarada: [5] documenta instrumentos, operaciones de procesamiento, parámetros, flujos de trabajo y procedencia; no afirma directamente que el archivo disponible más temprano esté necesariamente transformado.

Lo importante es documentar el estado más temprano que se conserva y las transformaciones entre estados.

En qué punto estamos

Los llamados datos brutos pueden incorporar ya calibración, filtrado, digitalización u otras transformaciones realizadas por instrumentos y sistemas de adquisición. «Bruto» es por tanto una etiqueta relativa al flujo de trabajo, no garantía de realidad intacta. Esto vuelve más importante la procedencia. La siguiente pregunta es qué ocurre cuando los archivos se preservan pero desaparece el conocimiento contextual necesario para interpretarlos.

11. Los datos pueden conservarse y, sin embargo, volverse inutilizables

Un archivo puede permanecer intacto byte a byte mientras desaparecen el software, el esquema, el diccionario de datos, las unidades o la procedencia necesarios para interpretarlo. NIST señala que unos metadatos deficientes pueden volver inutilizables conjuntos de datos importantes cuando sus creadores ya no están disponibles. [5]

C11. La preservación de los bytes no equivale a la preservación del significado ni de la reutilizabilidad. [5]

Por eso, los registros a largo plazo dependen de documentación, formatos accesibles, identificadores y metadatos contextuales, no solo de almacenamiento.

En qué punto estamos

El almacenamiento a largo plazo puede conservar bytes mientras se pierde el conocimiento necesario para interpretar variables, decisiones de procesamiento o contexto instrumental. Preservar es necesario pero no suficiente para reutilizar. El límite epistémico es semántico: supervivencia del archivo no implica supervivencia del significado. La siguiente pregunta es cuánto debe reconstruirse de la cadena observación-análisis para sostener reproducibilidad.

12. Los registros favorecen la reproducibilidad solo cuando la cadena puede reconstruirse

La reproducibilidad exige más que conservar los números finales. El tratamiento de la reproducibilidad por las National Academies subraya el acceso a los datos y a la información computacional necesaria para volver a calcular resultados, mientras que NIST vincula la reproducibilidad con metadatos, código, métodos e instrumentos documentados. [5][6]

C12. Los registros contribuyen a la reproducibilidad cuando conservan suficiente información sobre datos, metadatos, código, métodos y procedencia para reconstruir el resultado relevante. [5][6]

Esto no promete que todo estudio pueda recrearse exactamente. Algunas observaciones son únicas, destructivas o históricamente irrepetibles. La función del registro es conservar, hasta donde permita el contexto de investigación, el trayecto reconstruible desde la observación hasta la afirmación.

En qué punto estamos

Los registros apoyan reproducibilidad cuando otro investigador puede reconstruir la cadena relevante de datos, código, transformaciones, versiones y contexto para repetir el recorrido computacional. La reconstructibilidad no garantiza una interpretación científica idéntica. La siguiente pregunta es práctica: ¿qué auditoría mínima debe realizarse antes de tratar un conjunto de datos o una tabla como registro evidencial duradero?

13. Una auditoría práctica del registro

Para un conjunto de datos importante, pregunte:

1. ¿Qué representa una fila o registro?
2. ¿Qué significa cada variable y cuáles son sus unidades?
3. ¿Cómo se distinguen los valores ausentes, censurados y no aplicables?
4. ¿Qué registro es original, cuál está depurado y cuál es derivado?
5. ¿Qué transformaciones los conectan?
6. ¿Qué instrumento, método, software y parámetros produjeron los valores?
7. ¿Puede identificarse la versión exacta usada para una afirmación?
8. ¿Puede otra persona interpretar los datos sin depender de la memoria de quien los creó?

Esta lista es una síntesis editorial para este artículo, no un estándar universal de gestión de datos.

En qué punto estamos

Una auditoría práctica comprueba estructura, unidades, ausencia, metadatos, procedencia, versiones, identificadores y vínculos entre salidas transformadas y entradas. Superarla mejora la inspeccionabilidad, pero no demuestra que las mediciones subyacentes sean correctas. La siguiente pregunta es de frontera: ¿qué temas KNO vecinos poseen visualización, síntesis de evidencia y reproducibilidad más allá del problema más estrecho de los registros?

14. Frontera de este artículo

KNO-019 trata del papel epistémico de los registros y los datos estructurados. No intenta prescribir un esquema universal de bases de datos, una política de repositorio ni un flujo estadístico.

KNO-018 trató las representaciones visuales; KNO-022 tratará la computación como memoria externa; KNO-024 tratará la pérdida y el sesgo de preservación. Aquí la pregunta más estrecha es: **¿qué debe sobrevivir en un registro para que el razonamiento posterior siga siendo posible?**

En qué punto estamos

KNO-019 posee el problema epistémico de registros estructurados, metadatos, procedencia y objetos de datos versionados; no sustituye a KNO-018 sobre visualización ni a KNO-016 sobre síntesis de evidencia, y solo prepara KNO-021 sobre reproducibilidad abierta. Fijada esa frontera, el artículo puede resumir qué preservan los registros, qué no pueden preservar por sí solos y qué cambiaría nuestra confianza en ellos.

Lo que sabemos y cómo lo sabemos

Sabemos que una organización tabular explícita puede facilitar el análisis de variables y observaciones [1], que las convenciones de hoja de cálculo pueden reducir ambigüedades evitables y la mezcla de contenido original y derivado [2], que la procedencia puede representar entidades, actividades y agentes implicados en la producción de datos [3], y que la citación de datos exige suficiente especificidad para identificar el objeto o la versión citados [4]. NIST RDaF Version 2.0 trata explícitamente los metadatos, la procedencia, el versionado, los instrumentos, las operaciones de procesamiento y la reproducibilidad como partes de la gestión de datos de investigación [5]. Las National Academies vinculan asimismo la reproducibilidad con el acceso a los datos y a la información computacional necesarios para volver a calcular resultados [6].

Lo que todavía no sabemos

Ninguna estructura universal de registro garantiza la interpretabilidad futura. Distintas disciplinas requieren diferentes unidades, metadatos, granularidad de procedencia, periodos de conservación y formas de representar datos ausentes. Tampoco la documentación por sí sola establece que las mediciones sean válidas o que el análisis responda a la pregunta científica correcta.

Qué podría hacernos cambiar de opinión

Las conclusiones metodológicas de este artículo tendrían que revisarse si apareciera evidencia de que estas distinciones fallan de forma sistemática en dominios importantes, si modelos de procedencia o versionado más eficaces sustituyeran las prácticas actuales, o si futuras

infraestructuras hicieran recuperables por otros medios algunos registros contextuales que hoy son necesarios. Los cambios de herramienta no eliminarían el requisito de fondo: que un lector pueda identificar e interpretar el objeto probatorio realmente utilizado.

Si solo recuerdas una cosa

Los datos conservan conocimiento solo cuando el registro preserva suficiente significado, procedencia y contexto de versión para que los valores puedan interpretarse.

Artículos relacionados

- KNO-016 · *Cómo se construye la evidencia: integrando las herramientas*
- KNO-018 · *Imágenes, diagramas y visualización*
- KNO-022 · *La computación como memoria externa*
- KNO-024 · *Conocimiento perdido y sesgo de preservación*

Referencias

- [1] Hadley Wickham. “Tidy Data.” *Journal of Statistical Software* 59(10), 2014. DOI: 10.18637/jss.v059.i10.
- [2] Karl W. Broman; Kara H. Woo. “Data Organization in Spreadsheets.” *The American Statistician* 72(1):2–10, 2018. DOI: 10.1080/00031305.2017.1375989.
- [3] World Wide Web Consortium (W3C). *PROV-Overview: An Overview of the PROV Family of Documents*. W3C Working Group Note, 30 de abril de 2013. URI estable de la nota del W3C: <https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/>
- [4] Data Citation Synthesis Group. *Joint Declaration of Data Citation Principles*. FORCE11, 2014. Editora: Maryann E. Martone. DOI: 10.25490/a97f-egyk.
- [5] Robert Hanisch; Debra L. Kaiser; Alda Yuan; Andrea Medina-Smith; Bonnie C Carroll; Eva Campo. *NIST Research Data Framework (RDaF): Version 2.0*. NIST Special Publication 1500-18r2, 7 de febrero de 2024. DOI: 10.6028/NIST.SP.1500-18r2.
- [6] National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press, 2019. DOI: 10.17226/25303.
-