

KNO-019 · Data, Tables and Records

KNO · Ways of Knowing

Juan Segurola

v0.7 · EN

All Rights Reserved

Article Abstract

Scientific work leaves behind tables, spreadsheets, datasets and archives, but the survival of a file does not by itself tell us what later investigators can learn from it. What must a record preserve if someone years later is to interpret an observation responsibly and independently, distinguish absence from non-detection, identify the version that was analysed and reconstruct transformations between acquisition and publication? This article examines those questions across tabular structure, spreadsheet practice, metadata, provenance, versioning, persistent identifiers and the ambiguous category of “raw data.” It asks where simple storage ends and an auditable scientific record begins, how much context must travel with numerical values, and which parts of a data-producing history need to remain visible for reanalysis, correction and long-term reuse.

In 60 Seconds

Data become durable evidence only when their structure and history remain interpretable. Tables make some relationships easy to inspect, but spreadsheet convenience can also hide units, types, formulas or missing-value conventions. Metadata preserve the context needed to understand what values mean. Provenance records how data were produced or transformed. Versioning matters because a citation to “the dataset” is ambiguous if the underlying object changes, while persistent identifiers improve reference stability without replacing documentation. Missing values must be represented explicitly because absence, non-detection and not-applicable are not the same state. “Raw data” are also not automatically untouched: instruments and acquisition systems may already have applied calibration or processing. Preservation therefore succeeds only when future users can reconstruct enough of the chain from observation through processing to analysis. A good record is not merely a file that survives; it is a file whose meaning, lineage and version remain auditable by later investigators.

Why This Question Matters

Modern science depends on records that outlive the moment of observation. Measurements are copied into tables, transformed by software, merged with metadata, deposited in repositories and cited by later work. Each step can preserve information or quietly destroy context. A dataset may be numerically intact yet unusable because variable definitions, units, exclusions or processing history are missing. The problem matters for reproducibility, reanalysis, error correction and long-term preservation. Understanding records as structured, versioned and provenance-bearing objects helps distinguish “data still exist” from the stronger claim that another person can still understand what those data mean and how they became evidence.

1. A record is not the event

A record is a representation left by an event, observation, measurement or decision. It may preserve numbers, text, images, timestamps or categories, but it is not identical to what produced them. **C01. Records extend memory by making observations persistent, but persistence does not guarantee interpretability. Declared synthesis:** [3] and [5] support provenance, metadata and reconstruction; they do not directly establish the interpretability clause.

This distinction matters because later users usually encounter the record after the original context has disappeared. The epistemic task is therefore not only to store content, but to retain enough information about its production that another person can understand what the content means.

Where That Leaves Us

A record is a representation of an event or observation, not the event itself, so preservation always depends on what was encoded and what was omitted. That claim is conceptual rather than a denial that records can be reliable evidence. The next question is therefore structural: what organization must data have before relationships among observations can be inspected consistently?

2. Data need structure

A collection of values becomes easier to inspect and analyse when its structure is explicit. Wickham's tidy-data formulation is one influential convention: variables occupy columns, observations occupy rows, and different kinds of observational units occupy different tables. [1]

C02. A consistent tabular structure can reduce ambiguity and make transformation, modelling and visualisation easier, but it is a convention for organizing data rather than a universal ontology of the world. [1]

A row is not naturally an “observation” until someone defines the observational unit. The same event can be represented at the level of a person, visit, instrument reading, image, trial or transaction. The table therefore embodies a modelling decision.

Where That Leaves Us

Structured data reduce ambiguity by making variables, observations and values explicit, but no table design is neutral with respect to later operations. The epistemic gain is consistency and inspectability, not automatic correctness. Once structure is explicit, the next question is what tables reveal well and what they can obscure when relationships are compressed into rows and columns.

3. Tables expose relationships and hide others

Tables are powerful because they align comparable values. They allow sorting, filtering, joining, grouping and calculation. Yet a table can conceal the history that produced its cells. A final column may contain a value generated through several transformations whose intermediate steps are absent.

C03. Tabular organization supports comparison and computation, but the table alone need not preserve the provenance of each value. Declared synthesis: [2] supports disciplined tabular organization and [3] supports provenance; neither source directly states the non-preservation clause.

For this reason, a clean table and an auditable record are not the same thing. Good organization helps analysis; provenance helps reconstruction.

Where That Leaves Us

Tables expose comparisons, joins and patterns efficiently while suppressing other relationships that may require different representations. Their usefulness depends on matching structure to the question being asked. That limitation leads to a practical danger: when table-like data are handled in flexible spreadsheets, what conveniences can quietly weaken the integrity and auditability of the record?

4. Spreadsheet convenience can create epistemic fragility

Spreadsheets are useful because they combine data entry, calculation and presentation. The same flexibility can also mix data with formatting, comments, multiple variables in one cell, embedded calculations or inconsistent missing-value codes. Broman and Woo recommend simple rectangular organization, one piece of information per cell, consistent missing-value conventions and separation of raw data from calculations. [2]

C04. Spreadsheet practices that mix raw observations, formatting and derived calculations can make later interpretation and analysis more error-prone. [2]

This is not an argument against spreadsheets. It is an argument for treating the record structure as part of the method.

Where That Leaves Us

Spreadsheets are powerful working environments, but mixed types, hidden formulas, manual formatting and overloaded cells can make data fragile or ambiguous. The risk is not that spreadsheets are inherently invalid; it is that convenience can hide structure. The next question is what information must accompany the values themselves so that later users can still interpret the record correctly.

5. Metadata preserve meaning around values

Metadata describe the data. They may identify variables, units, instruments, dates, locations, creators, methods, software, access conditions or links to related records. NIST's Research Data Framework treats metadata as essential to effective use, reuse and preservation over time. [5]

C05. Values become more reusable when metadata preserve the definitions and conditions needed to interpret them. [5]

A column labelled temperature is incomplete if the scale, sensor, location, timing and processing are material to the question. Metadata do not magically remove uncertainty; they preserve context needed to reason about it.

Where That Leaves Us

Metadata preserve the context around values: variable definitions, units, methods, identifiers and other information needed to interpret what a record means. Complete metadata do not guarantee valid science, but missing context can make valid observations unusable. The next question is historical rather than descriptive: how can a record preserve not only what values are, but how they came to exist?

6. Provenance records how data came to be

The W3C PROV family defines provenance in terms of entities, activities and agents involved in producing a piece of data or other object. It is intended to support interoperable descriptions of origins and transformations. [3]

C06. Provenance records can document where data came from, who or what acted on them, and which processes contributed to the resulting object. [3]

Provenance is therefore different from a descriptive label. A dataset may have excellent variable names yet still fail to show which source file, instrument run, script or transformation produced the version being analysed.

Where That Leaves Us

Provenance records the lineage of data through acquisition, transformation and use, allowing later users to inspect how an object was produced. Provenance improves reconstructability without proving that every transformation was appropriate. Once lineage matters, a further problem appears: if the dataset changes over time, which exact state of that evolving object does a later citation identify?

7. Versioning changes what a citation points to

Data can change after initial creation: errors are corrected, records are added, classifications are revised and processing pipelines are updated. If a claim depends on a particular state of a dataset, citing only the general dataset name can be insufficient.

The Joint Declaration of Data Citation Principles emphasizes specificity and verifiability, including information sufficient to identify a particular version, timeslice or granular portion of data. [4]

C07. When data change over time, version information can be necessary to determine which exact data supported a claim. [4][5]

The knowledge claim is tied not only to a dataset family but to an identifiable manifestation of it.

Where That Leaves Us

Versioning distinguishes successive states of a data object, so a reference can point to what was actually analyzed rather than to an unspecified moving target. A version label helps only if the version is preserved and identifiable. The next question is how persistent identifiers contribute to this problem and, equally important, which documentation problems they cannot solve on their own.

8. Persistent identifiers solve one problem, not every problem

Persistent identifiers help keep references stable when objects move. Data-citation principles recommend globally unique, persistent, machine-actionable identification. [4]

C08. A persistent identifier improves referential stability, but it does not by itself establish data quality, correct interpretation or evidential relevance. [4]

A DOI can tell us which object was cited. It cannot tell us whether the measurements were valid or whether the analysis answered the intended question. Identification and evaluation are separate epistemic tasks.

Where That Leaves Us

Persistent identifiers make references more stable and machine-resolvable, but they do not supply missing metadata, provenance or version semantics by themselves. They solve identification, not the whole problem of interpretability. The next question moves inside the table again: when a cell has no ordinary value, how should the record distinguish missingness from zero, non-detection or inapplicability?

9. Missing values are records too

Absence in a table can mean many things: not measured, not applicable, below detection, lost, refused, impossible, corrupted or truly zero. Treating these as interchangeable can change analysis.

C09. Missingness requires explicit representation because an empty field, a zero and an unobserved value can have different meanings. [2]

The exact coding scheme is domain dependent. The general methodological point is that a record should distinguish different states when those distinctions matter to interpretation.

Where That Leaves Us

Missing values are themselves states that require explicit representation, because different reasons for absence can imply different analyses and interpretations. Encoding missingness carefully does not recover information that was never observed, but it prevents several states from being collapsed into one. The next question is whether the familiar label “raw data” reliably identifies an untouched starting point.

10. Raw data are not always untouched data

The phrase “raw data” often suggests a pristine layer before interpretation. In practice, instruments may already apply calibration, filtering, compression, thresholding or proprietary conversion before a researcher receives a file. NIST's RDaF explicitly treats instruments, processing operations, parameters, workflows and provenance as relevant parts of data management. [5]

C10. The earliest available digital record may already contain processing decisions, so “raw” should not be assumed to mean untransformed. Declared synthesis: [5] documents instruments, processing operations, parameters, workflows and provenance; it does not directly state that the earliest available file is necessarily transformed.

What matters is to document the earliest retained state and the transformations between states.

Where That Leaves Us

So-called raw data may already include calibration, filtering, digitization or other transformations performed by instruments and acquisition systems. “Raw” is therefore a workflow-relative label rather than a guarantee of untouched reality. That makes provenance more important, not less. The next question is what happens when files are preserved successfully but the contextual knowledge required to interpret them disappears.

11. Data can be preserved yet become unusable

A file may remain byte-for-byte intact while the software, schema, codebook, units or provenance needed to interpret it disappear. NIST notes that poor-quality metadata can render important datasets unusable when their creators are no longer available. [5]

C11. Preservation of bytes is not equivalent to preservation of meaning or reusability. [5]

Long-term records therefore depend on documentation, accessible formats, identifiers and contextual metadata, not storage alone.

Where That Leaves Us

Long-term storage can preserve bytes while losing the knowledge needed to interpret variables, processing choices or instrument context. Preservation is therefore necessary but not sufficient for reuse. The epistemic limit is semantic: survival of a file does not ensure survival of meaning. The next question is how much of the observation-to-analysis chain must remain reconstructable for records to support reproducibility.

12. Records support reproducibility only when the chain is reconstructable

Reproducibility depends on more than retaining final numbers. The National Academies' treatment of reproducibility emphasizes access to data and computational information needed to recompute results, while NIST connects reproducibility with documented metadata, code, methods and instruments. [5][6]

C12. Records contribute to reproducibility when they preserve enough of the data, metadata, code, methods and provenance for the relevant result to be reconstructed. [5][6]

This is not a promise that every study can be exactly recreated. Some observations are unique, destructive or historically unrepeatable. The record's role is to preserve the reconstructable path from observation to claim as far as the research context permits.

Where That Leaves Us

Records support reproducibility when another researcher can reconstruct the relevant chain of data, code, transformations, versions and context well enough to repeat the computational path. Reconstructability still does not guarantee identical scientific interpretation. The next question is practical: what minimum audit should a researcher perform before treating a dataset or table as a durable evidential record?

13. A practical record audit

For an important dataset, ask:

1. What does one row or record represent?
2. What does each variable mean, and what are its units?
3. How are missing, censored and not-applicable values distinguished?
4. Which record is original, which is cleaned, and which is derived?
5. What transformations connect them?
6. Which instrument, method, software and parameters produced the values?
7. Is the exact version used for a claim identifiable?
8. Can another person interpret the data without relying on the creator's memory?

This checklist is an editorial synthesis for this article, not a universal data-management standard.

Where That Leaves Us

A practical record audit checks structure, units, missingness, metadata, provenance, versions, identifiers and the links between transformed outputs and their inputs. Passing such an audit improves inspectability but does not prove the underlying measurements are correct. The next question is one of boundary: which neighboring KNO topics own visualization, evidence synthesis and reproducibility beyond the narrower problem of records?

14. Boundary of this article

KNO-019 concerns the epistemic role of records and structured data. It does not attempt to prescribe a universal database schema, repository policy or statistical workflow. KNO-018 addressed visual representations; KNO-022 will address computation as external memory; KNO-024 will address loss and preservation bias. Here the narrower question is: **what must survive in a record for later reasoning to remain possible?**

Where That Leaves Us

KNO-019 owns the epistemic problem of structured records, metadata, provenance and versioned data objects; it does not replace KNO-018 on visualization or KNO-016 on evidence synthesis, and it only prepares the ground for KNO-021 on open reproducibility. With that boundary fixed, the article can now summarize what records can preserve, what they cannot, and what would change our confidence in them.

What We Know, and How

We know that explicit tabular organization can make variables and observations easier to analyse [1], that spreadsheet conventions can reduce avoidable ambiguity and mixing of raw and derived content [2], that provenance can represent entities, activities and agents involved in producing data [3], and that data citation requires specificity sufficient to identify the object or version being cited [4]. NIST RDaF Version 2.0 explicitly treats metadata, provenance, versioning, instruments, processing operations and reproducibility as parts of research-data management [5]. The National Academies likewise connect reproducibility with access to the data and computational information needed to recompute results [6].

What We Still Don't Know

No universal record structure guarantees future interpretability. Different disciplines require different units, metadata, provenance granularity, retention periods and representations of missingness. Nor can documentation alone establish that measurements were valid or that an analysis answered the right scientific question.

What Could Change Our Minds

The methodological conclusions here would need revision if evidence showed that the proposed distinctions systematically fail in important domains, if more effective provenance or versioning models replace current practice, or if future infrastructures make some currently necessary contextual records reliably recoverable by other means. Changes in tools would not remove the underlying requirement that a reader be able to identify and interpret the evidential object actually used.

If You Remember Only One Thing

Data preserve knowledge only when the record preserves enough meaning, provenance and version context to make the values interpretable.

Related Articles

- KNO-016 · *How Evidence Is Built: Putting the Tools Together*
- KNO-018 · *Images, Diagrams and Visualisation*
- KNO-022 · *Computation as External Memory*
- KNO-024 · *Lost Knowledge and Preservation Bias*

References

- [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 April 2013. Stable URI for the W3C Note: <https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/>
- [4] Data Citation Synthesis Group. *Joint Declaration of Data Citation Principles*. FORCE11, 2014. Editor: 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 February 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.