

KNO-024 · Lost Knowledge and Preservation Bias

KNO · Ways of Knowing

Juan Segurola

v0.3 · EN

All Rights Reserved

Article Abstract

What survives from the past is not a random sample of what once existed. Bones weather, manuscripts burn, institutions discard records, file formats become unreadable and datasets vanish when storage devices or contact addresses fail. The surviving archive is therefore shaped by processes of preservation, selection and destruction before any historian or scientist begins to interpret it. Archaeology formalized this problem through formation processes that separate past activity from the material record recovered later.[1] Paleontology likewise studies how preservation filters organisms and environments.[2,3] Archives add human decisions about what is created, kept, described and made accessible.[4] Digital information removes some physical constraints but introduces dependence on formats, software, migration and active stewardship.[5,6] This article asks how knowledge claims should change when absence may reflect non-preservation rather than non-existence, and how preservation systems can make their own selection effects more visible.

In 60 Seconds

Records survive through filters. In archaeology, objects move from active use into archaeological context through cultural and natural formation processes, so what is excavated is already transformed.[1] In paleontology, weathering, transport, burial and mineralization make some organisms and environments much more likely to enter the fossil record than others.[2,3] Archives introduce another layer: people and institutions decide what to document, retain, classify and permit others to access.[4] Digital media do not abolish loss. Vines and colleagues found that the availability of research datasets declined strongly with article age in the sample they studied, illustrating how fragile data become when preservation depends on individual custodians.[5] Digital preservation therefore requires migration, integrity checking, metadata and institutional continuity rather than mere copying.[6] The methodological lesson is simple but demanding: absence in a surviving record is evidence only after the processes that could have removed, hidden or never recorded the item have been considered.

Why This Question Matters

We reason from survivors constantly: fossils, laboratory notebooks, administrative files, archaeological sites, photographs, datasets and source code. Because these survivors are visible, they can feel like the past itself. They are not. Every record has a survival history, and that history can be correlated with the property we want to infer. Large bones preserve differently from fragile tissue; official archives preserve differently from informal speech; maintained repositories preserve differently from forgotten drives. Preservation bias matters whenever the mechanism of survival can distort the apparent frequency, diversity or chronology of what we observe.

1. The Record Is Produced by Survival

The Question If most traces disappear before anyone can inspect them, what does the surviving record actually sample? The answer depends not only on what happened in the past, but also on which traces were created, preserved, collected, described and made discoverable to later observers.

A past event and a surviving trace are not the same object. Between them lie processes that create, transform, move, preserve and destroy records. Schiffer's distinction between systemic and archaeological context made formation processes central to archaeological inference: objects recovered later have post-use histories that affect their location and association.[1] Before treating a trace as a sample of past reality, one must ask what made that trace likely to survive.

Claim C01. Archaeological evidence is shaped by formation processes between active use and later recovery. [1]

Where That Leaves Us

The surviving record is generated twice: first by the events that create traces, then by the processes that preserve some traces and remove others. Inference must model both. Preservation is therefore not a nuisance added after observation; it is part of the observation process itself. The next section shows how strongly this matters in the physical record of organisms and environments.

2. Taphonomy Filters the Biological Past

Behrensmeyer calibrated stages of bone weathering with known-age carcasses and showed that weathering depends on time and local conditions.[2] Kidwell and Flessa review variation in completeness, temporal resolution and ecological fidelity of fossil records.[3] Fossil abundance is therefore not identical to past abundance: hard parts, rapid burial and depositional setting can alter preservation probability. Taphonomy treats that filter as evidence rather than noise.

Claim C02. Bone weathering varies with time and environmental conditions and can inform taphonomic interpretation. [2]

Claim C03. The fossil record differs in completeness and temporal/ecological resolution across contexts. [3]

Where That Leaves Us

Fossils are filtered observations. Weathering, transport, burial and mineralization change which organisms, body parts and time intervals become visible. Taphonomy does not make the fossil record useless; it makes the preservation process part of the evidence. The same principle applies to human records, but with an additional complication: people and institutions intentionally decide what to document and keep.

3. Archives Are Built, Not Merely Inherited

Schwartz and Cook argue that archives and records participate in the construction of social memory rather than functioning as neutral containers of facts.[4] Decisions about creation, appraisal, description, access and custody shape what later users can encounter. A missing record can therefore mean non-occurrence, but it can also mean non-creation, destruction, exclusion from acquisition, poor description or restricted access.

Claim C04. Archives and records are shaped by human decisions about creation, appraisal, description and access. [4]

Where That Leaves Us

Human archives add intentional selection to physical survival. Records can disappear because no one created them, because institutions did not retain them, because description made them hard to find or because access remained restricted. Archival provenance therefore includes custody and policy. The next issue is asymmetry: official systems often preserve some voices and activities more systematically than informal ones.

4. Missingness Is Often Structured

Preservation bias is most dangerous when missingness correlates with the question being asked. Durable stone and metal survive differently from textiles, wood or speech; formal institutions document differently from informal communities; crises may leave denser paperwork than routine success. The key questions are which categories had opportunities to create traces, which traces were likely to survive and which were likely to be collected and discovered.

Claim C05 - Declared synthesis. Missingness can be structured when survival probability depends on the property being studied. **Declared synthesis:** [1-4] support selective formation, survival and archival processes; describing missingness as structured when survival depends on the studied property is the synthesis.

Claim C06 - Declared synthesis. Absence in an archive is not automatically evidence that an event did not occur. **Declared synthesis:** [1,4] support multiple pathways by which records can fail to form or survive; the inference that archival absence is therefore not automatically event-absence is the synthesis.

Where That Leaves Us

Preservation bias is strongest when the mechanism of loss is related to the phenomenon under study. Then the surviving record can exaggerate durable, official, exceptional or institutionally valued material. Correct interpretation requires a model of who could create records and which records could survive. Digital media seem at first to remove many physical filters, but they introduce their own dependencies and failure modes.

5. Digital Does Not Mean Permanent

Vines and colleagues requested datasets from 516 biological papers spanning two to twenty-two years and found that, among cases where authors reported data status, the odds of a dataset being extant fell by 17 percent per year in their sample.[5] Broken contact addresses and obsolete storage were major obstacles. The estimate is sample-specific, but it demonstrates that digital data can disappear rapidly without institutional preservation.

Claim C07. In the Vines et al. sample, data availability declined strongly with article age. [5]

Claim C08. Digital preservation requires active stewardship beyond simple copying. [5,6]

Where That Leaves Us

Digital information can be copied cheaply, yet persistence still requires active stewardship. Data may vanish through obsolete devices, broken links, missing credentials, format drift or loss of context. The Vines study shows one empirical form of this decay in research data. Preservation therefore shifts from protecting one physical object to maintaining an interpretable chain of files, software, metadata and institutions.

6. Digitization Is Not the Same as Digital Preservation

Conway distinguishes digitization for preservation from digital preservation as a sociotechnical activity.[6] Scanning can reduce handling of an original and widen access, but the scan itself becomes a digital object that requires management. Resolution, colour, compression, metadata and file format determine what the surrogate captures, while migration and validation determine whether that surrogate remains usable.

Claim C09. Digitization and digital preservation are distinct activities. [6]

Claim C10 - Declared synthesis. A digital surrogate preserves only what its capture and metadata choices encode. **Declared synthesis:** [6] supports distinguishing digitization choices from digital preservation; the clause that a surrogate preserves only what capture and metadata encode is the synthesis.

Where That Leaves Us

Digitization creates a new preservation object rather than ending preservation work. What is captured depends on conversion choices, and the resulting files require ongoing management. A digital surrogate can improve access while omitting features of the source or becoming obsolete itself. The next problem follows: preservation systems must make their own selections and transformations visible enough for later users to judge what survived.

7. Preservation Needs Provenance of Loss and Change

A useful archive can record not only what it holds but, where possible, what was transformed, migrated, deaccessioned or known to be missing. Checksums document byte integrity; migration logs record format changes; collection policies explain selection boundaries. These records cannot reconstruct everything that was lost, but they make some preservation processes auditable and help distinguish original evidence from later transformation.

Claim C11 - Declared synthesis. Preservation systems can improve auditability by recording migration, integrity and selection history. **Declared synthesis:** [6] supports the need to manage digital objects across preservation actions; the claim that recording migration, integrity and selection history improves auditability is the synthesis.

Where That Leaves Us

Preservation systems can reduce bias not only by saving more material, but by documenting the filters they apply. Integrity records, migration histories and collection policies expose parts of the survival process to later scrutiny. They cannot recover what was never recorded, yet they can prevent silent transformation from masquerading as original evidence. The final section asks how claims should be phrased when the preservation filter remains only partly known.

8. Absence Requires a Model of Preservation

An absence is strongest evidence when the system was likely to create, preserve and expose the relevant record had the event occurred. It is weaker when any of those probabilities are low or unknown. Archaeologists model formation processes, paleontologists study taphonomy, archivists examine appraisal and custody, and digital preservation studies formats and stewardship. Different domains express the same discipline: interpret the record through the process that made it observable.

Claim C12 - Declared synthesis. Absence is strongest evidence when creation, preservation and discoverability would all have been likely. **Declared synthesis:** [1-6] support distinct creation, survival, archival and digital-preservation filters; combining them into this criterion for when absence is strongest evidence is the synthesis.

Where That Leaves Us

Lost knowledge is not only missing content; it is uncertainty about the filtering process that produced what remains. Absence becomes informative when creation, preservation and discoverability would all have been likely. Where those conditions are weak, confidence should fall accordingly. KNO-024 closes the Ways of Knowing sequence by making preservation itself part of evidence: what survives constrains knowledge only through a model of how survival happened.

What We Know, and How

- Archaeological records are shaped by formation processes between use and recovery.[1]
- Bone weathering and other taphonomic processes affect which biological traces survive and how they are interpreted.[2,3]
- Archives embody decisions about creation, appraisal, description and access rather than functioning as neutral containers.[4]
- Research datasets can become unavailable over time when preservation depends on individual custodians.[5]

What We Still Don't Know

- How much undocumented material has already been lost from contemporary digital research ecosystems.
- How preservation probabilities vary across underrepresented communities and informal record-keeping systems.
- Which digital preservation strategies will remain interpretable over century-scale technological change.
- How best to quantify preservation bias when creation, loss and discovery probabilities are all uncertain.

What Could Change Our Minds

- Empirical records showing that major classes of supposedly fragile evidence preserve at the same rate as durable evidence across relevant contexts.
- Archival studies demonstrating that appraisal and access decisions do not measurably affect later historical inference in comparable collections.

- Longitudinal digital studies showing negligible loss of research data without institutional stewardship.
- New preservation methods that make format, software and metadata obsolescence effectively irrelevant over long intervals.

If You Remember Only One Thing

A surviving record is evidence about the past and about the processes that selected what could survive; absence is interpretable only after both are considered.

Related Articles

- KNO-009 · Archives, Libraries and Collective Memory
- KNO-019 · Data, Tables and Records
- KNO-021 · Open Science and Reproducibility

References

[1] Michael B. Schiffer. Archaeological Context and Systemic Context. *American Antiquity* 37(2), 156–165 (1972). DOI 10.2307/278203. <https://doi.org/10.2307/278203> [2] Anna K. Behrensmeyer. Taphonomic and ecologic information from bone weathering. *Paleobiology* 4(2), 150–162 (1978). DOI 10.1017/S0094837300005820. <https://doi.org/10.1017/S0094837300005820> [3] Susan M. Kidwell; Karl W. Flessa. The quality of the fossil record: populations, species, and communities. *Annual Review of Earth and Planetary Sciences* 24, 433–464 (1996). DOI 10.1146/annurev.earth.24.1.433. <https://doi.org/10.1146/annurev.earth.24.1.433> [4] Joan M. Schwartz; Terry Cook. Archives, Records, and Power: The Making of Modern Memory. *Archival Science* 2(1–2), 1–19 (2002). DOI 10.1007/BF02435628. <https://doi.org/10.1007/BF02435628> [5] Timothy H. Vines; Arianne Y.K. Albert; Rose L. Andrew; Florence Débarre; Dan G. Bock; Michelle T. Franklin; Kimberly J. Gilbert; Jean-Sébastien Moore; Sébastien Renaut; Diana J. Rennison. The availability of research data declines rapidly with article age. *Current Biology* 24(1), 94–97 (2014). DOI 10.1016/j.cub.2013.11.014. <https://doi.org/10.1016/j.cub.2013.11.014> [6] Paul Conway. Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas. *The Library Quarterly* 80(1), 61–79 (2010). DOI 10.1086/648463. <https://doi.org/10.1086/648463>