

Methods note · Observation and data

Cite the release, not the latest link

Theo Nguyen · 6 October 2026

Fictional author and institute · Editorial demonstration, not peer reviewed

ABSTRACT

A persistent dataset address can resolve successfully while no longer providing the evidence used in a published calculation. This methods note compares FORCE11 citation principles, DataCite versioning guidance, and Research Data Alliance recommendations for evolving data. A fictional index revision separates corrected observations from expanded coverage and demonstrates why collection identifiers, release identifiers, subsets, and analysis outputs answer different questions. The discussion examines immutable releases, timestamped queries, stable sorting, checksums, and persistent landing pages without presenting any one repository design as mandatory. It distinguishes exact reproduction from scientific robustness and explains why reproducibility does not override legitimate restrictions. The resulting recommendation is a compact citation backed by recoverable authorised evidence and an inspectable method, with explicit limits when retrieval becomes impossible.

Keywords: data citation · dataset versions · reproducibility · provenance

A source-based editorial text, not a report of original Still North research. Examples and imagery are illustrative.

Two readers, one address, different evidence

A dataset address can remain stable while the evidence behind it changes. Late observations arrive, quality flags are revised, coordinates are corrected, and processing improves. These changes are often desirable. The difficulty begins when a paper cites only the address of the current download, leaving a later reader unable to recover the release that supported the published calculation. A persistent link is necessary infrastructure, but persistence of an address is not the same as fixity of its contents.

This methods note compares three primary sources of citation guidance: the FORCE11 Joint Declaration, DataCite's versioning documentation, and the Research Data Alliance recommendations for evolving data. The comparison asks what each contributes to a reproducible citation and where an archive must make an explicit implementation choice. It is not a survey of repository performance. The examples are invented, and no identifiers shown here are represented as registered dataset DOIs.

Identify the object before choosing the identifier

A collection, a release, a subset, and an analysis output are related objects, not synonyms. A collection can describe an ongoing observation programme. A release fixes one published state of that programme's records. A subset selects particular variables, sites, dates, and quality categories from that state. An output applies a method to the subset. Citing the collection gives credit and context; it may not identify the exact evidence used.

FORCE11's principles make this distinction substantive. Specificity and verifiability require information sufficient to identify the relevant time slice, version, or granular portion. Persistence also extends to metadata about a dataset's disposition after the data themselves cease to

be available. A citation therefore has both an evidential function and a stewardship dependency: an identifier cannot preserve a release that nobody has retained.

Before writing a citation, establish what the repository promises. Does the identifier resolve to a fixed release or the whole evolving collection? Can an older release still be retrieved? Does a minor update replace files? What happens after a withdrawal? The answers belong in the repository's documented policy. Guessing that every DOI denotes immutable bytes creates a false assurance that the identifier system does not itself provide.

Fictional case: a revised seasonal index

Consider an invented observation archive. Release 1.0 contains 90 daily values whose sum is 900 index units. A paper reports their arithmetic mean as 10.0. Release 2.0 corrects ten affected values by adding two units to each and incorporates ten late values averaging 20.0. The revised sum is 1,120 over 100 days, giving 11.2. These are transparent teaching numbers, not environmental observations or a realistic error distribution.

A reader downloading only the latest file cannot reproduce 10.0, even with exactly the original averaging method. The discrepancy has two causes: correction of existing evidence and expansion of coverage. A change log saying merely "updated dataset" would conceal that distinction. The release note should identify the affected records and the nature of both changes at an appropriate level, without disclosing protected information.

The evidence record for the original paper would retain release 1.0, its selected input, and the calculation method where permission allows. A new analysis could cite release 2.0 and explain why 11.2 is now preferred. Reproducing a superseded result is not endorsing it. It enables an auditor to distinguish an original computational error from a valid calculation made with an earlier evidential state.

Object	Information needed	Question answered
Collection record	Programme description, creator credit, and links to releases	Where does this evidence belong?
Release 1.0	Version, fixed files, release date, and provenance	Which published state supported 10.0?
Analysis subset	Variable, date boundaries, quality rule, and stable record selection	Which records entered the calculation?
Mean calculation	Weighting, missing-data treatment, and method version	How did those records become the reported result?

Version labels need a declared meaning

DataCite recommends updating metadata for minor content changes and assigning a new DOI for major changes, with related identifiers linking the versions. It leaves the major-versus-minor distinction to stewards. This flexibility is important, but it means that a version number alone cannot tell an analyst whether a change affects their conclusion. A one-cell correction can matter more to a particular claim than a large addition elsewhere.

The fictional archive could choose to preserve immutable analytical releases even where its repository would permit replacement. That is an archive policy, not a universal DOI requirement. It should distinguish changes to descriptive metadata from changes to scientific values, quality flags, units, coordinate interpretation, and processing. Release notes should say what changed and why, while the version relations say how the objects connect.

A collection identifier can provide the continuing home for a programme, while release-specific identifiers support exact citations. DataCite describes relations such as `IsVersionOf` and `IsNewVersionOf` for these purposes. The public citation should point to the specific release used, with the collection linked for discovery where useful. It should not make readers infer the analytical version from the date on an unrelated web page.

Dynamic queries require historical state

A live database complicates matters because an analyst may use a query rather than a whole release. The RDA Working Group's 2015 recommendations propose versioned, timestamped data, stored queries, stable sorting, result checksums, and persistent identification of the query and its context. This is a precise way to identify evolving subsets, provided that the infrastructure can recreate the underlying state.

Saving the text “select accepted observations for October” is insufficient. Accepted status might change later. The query's date boundaries need defined inclusivity, its timestamps need an unambiguous time basis, and its source state must be recoverable. A current query with an old retrieval date in a caption does not restore records overwritten since retrieval. Distributed sources may also have different update times, making one wall-clock timestamp an inadequate description of the combined evidence.

A checksum offers a different assurance. It can establish that a retrieved file matches a previously recorded byte sequence, subject to the chosen method. It cannot explain a variable, identify who authorised sharing, or prove the scientific correctness of a value. Byte differences may also reflect harmless changes in compression or row ordering.

Stable sorting and documented serialisation are therefore useful when checksums are intended to verify query results rather than arbitrary exported files.

Keep the citation compact and the evidence package complete

A useful human-readable citation identifies the creators, dataset title, publisher or repository, release year, version, and persistent identifier in the relevant scholarly style. The landing page can hold detailed provenance, file inventories, fixity information, and version relations. The analysis record then identifies the subset and method. Making a citation enormous does not compensate for missing infrastructure behind its link.

For small archives, preserving a fixed, documented release and the exact authorised analysis input may be simpler than operating a historical query service. The RDA approach solves a substantial dynamic-data problem, but it should not be imitated incompletely. An archive that stores queries without preserving data history has gained complexity without the central reproducibility benefit. Choose the simpler arrangement that can actually honour the evidential promise.

Responsibility should survive a staff change. Someone must maintain identifier resolution, explain corrections, preserve permitted older states, and record when access conditions change. If data must be removed, a persistent landing page can explain their disposition without retaining sensitive payloads publicly. Both `FORCE11` and `RDA` leave room for persistent metadata when retrieval is no longer possible. Reproducibility is an important commitment, not a justification for overriding legitimate restrictions.

Limits and implications

This note does not prescribe a universal version-number scheme or guarantee that a particular repository will preserve every release. Exact rerunning may remain impossible when software, licensed inputs, permissions, or external services cannot be preserved. Numerical reproducibility also differs from scientific robustness: recovering 10.0 does not show that the original question, sampling, or model was appropriate.

The practical implication is to separate three promises: the evidence can be located, the specific evidential state can be recovered where authorised, and the transformation can be inspected. A versioned citation supports the first two; a preserved method supports the third. Together they let correction improve a record without erasing the route by which an earlier conclusion was reached.

References

1. Data Citation Synthesis Group, Martone, M., editor (2014). Joint Declaration of Data Citation Principles. FORCE11.
2. DataCite (documentation consulted 2026). Versioning.
3. Rauber, A., Asmi, A., van Uytvanck, D., and Pröll, S. (2015). Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation, revision of 20 October 2015.