



P. Veloso

Engineering Leader · Solutions Architect · Technical Lead · Researcher

[Home](#)

[Experience](#)

[Portfolio](#)

[News](#)

[Contact](#)

[↓ CV](#)



A Reproducible Research-Cataloguing Workflow

[Home](#) / [Portfolio](#) / **A Reproducible Research-Cataloguing Workflow**

INDEPENDENT RESEARCH INFRASTRUCTURE · 2026

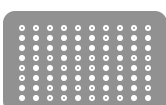
The problem

A serious literature project accumulates a hundred or so sources from many places: preprint servers, journals, library-discovery databases, direct downloads from authors. Left alone, that becomes an unsearchable folder of files with machine-generated names, no record of where each one came from, no protection against a lost or evicted download, and no way to reproduce the set later or prove to a reviewer that it is what it claims to be. I wanted a workflow where the source base is reproducible, resilient to the loss of the originals, and auditable all the way from a raw file to a catalogued record, so that the evidence base is a first-class research artefact rather than a convenience.

What I built

I designed a project-agnostic cataloguing standard and applied it to this project's source base. It rests on a small number of deliberate choices, each aimed at a specific failure mode.

A local copy of every source lives inside the project, so the archive survives deletion or eviction of the original download. The reference base does not depend on a link that may rot or a cloud file that may be removed later. This is the difference between citing a source and holding it.



© 2026 Pedro Veloso

[LinkedIn](#)

[GitHub](#)

[ORCID](#)

[Contact](#)

tent hash does not.

Every record has a stable identifier, assigned once and never reused, so a source can be referred to consistently across every document in the project even as the set grows and is re-sorted. When a later document needs to point at a source, it points at an identifier that will still mean the same thing months later.

Classification is driven by a small metadata file placed beside each source rather than by guesswork over the filename, so the class of each item is explicit, reviewable, and correctable without touching the source itself.

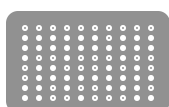
The catalogue exports to several formats at once: a queryable database, a spreadsheet, and human-readable text, so it can be used by a person reading, a program querying, or a reviewer checking, without any of them being the single point of truth.

The workflow around it

The discipline is as much in the process as in the tool. New files are triaged first by their abstract for relevance, then read in full before anything is catalogued, so nothing enters the archive unread and no source is ever cited on the strength of its abstract alone. Off-topic drops are not deleted; they are moved to a holding area and recorded by content hash, so that if the same file is ever dropped again it is recognised as already considered rather than analysed a second time. Exact duplicates are collapsed to a single copy by hash. The machine catalogue is deliberately kept separate from the authoritative, hand-checked bibliography, so each does the job it is good at: the catalogue carries structure, hashes and provenance for every file, and the bibliography carries verified titles, authors, venues and identifiers that a human has confirmed. A structural record and a citation record are different things, and conflating them is how unverified metadata ends up in a reference list.

Why it is built this way

The choices are not arbitrary; they follow the way provenance is understood in the data management literature, which distinguishes where a piece of data came from, how it was derived, and why it is there ([Cheney et al., 2009](#)). A content hash plus a stable identifier plus a local copy answers the where and the how for every source, so a catalogued record can always be traced back



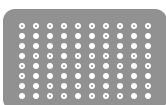
works ([Menotti et al., 2025](#)), which is the same instinct applied to assertions rather than files: record not just the claim but where it came from and how far it can be trusted.

The separation of the machine catalogue from the hand-checked bibliography reflects a methodological point as much as an engineering one. Design-science research treats the rigour of its evidence base as part of the contribution rather than an afterthought ([Hevner et al., 2004](#)), and its presentation and knowledge-contribution framework expects the design and its justification to be laid out so they can be assessed ([Gregor and Hevner, 2013](#)). Recent work on validity in design science makes the traceability of the evidence base an explicit criterion for stating and validating a knowledge claim ([Larsen et al., 2025](#)), and an analysis of knowledge-contribution paths frames how a general method and its instantiation should each be recorded ([Akoka et al., 2023](#)). An analysis of design-science doctoral theses found that scoping and evaluation are exactly where such projects most often fail ([Cater-Steel et al., 2019](#)), which is a direct argument for infrastructure that keeps the source base tight, explicit and checkable rather than sprawling and implicit. The workflow is built so a reviewer could reconstruct and re-check the source base from the retained copies and hashes, which is precisely the property those methods ask for.

Outcome, including a failure it survived

More than ninety sources were catalogued this way, each with a local copy, a content hash, a stable identifier, and explicit classification, and the whole set exports cleanly for reuse. The discipline paid off when the catalogue database later appeared to fail: every read returned zero bytes and the file reported as not a database. Because the raw files, their metadata sidecars, and the hand-checked bibliography were intact and independent of the database, the failure was contained to one derived artefact rather than the whole evidence base. The database turned out not to be corrupt at all but offloaded by the operating system's cloud-storage optimisation and left as a dehydrated stub, and reading the files through materialised them and brought it back whole. Nothing was lost, the derived exports were refreshed, and the offloading was recorded as a standing risk to the local-copy strategy with the setting to change so it does not recur. A system designed so that a failure in one part cannot take the others with it is what turned a frightening symptom into a routine recovery.

What it demonstrates



preference, treating resilience against data loss as a first-class goal, and separating the parts of a system so that a failure in one does not take the others with it.

Selected references

- Akoka, J., Comyn-Wattiau, I., Prat, N., Storey, V.C. (2023). Knowledge contributions in design science research: Paths of knowledge types. *Decision Support Systems*, 166. [DOI 10.1016/j.dss.2022.113898](https://doi.org/10.1016/j.dss.2022.113898).
- Cater-Steel, A., Toleman, M., Rajaeian, M.M. (2019). Design Science Research in Doctoral Projects: An Analysis of Australian Theses. *Journal of the Association for Information Systems*, 20(12), 1844-1869. [DOI 10.17705/1jais.00587](https://doi.org/10.17705/1jais.00587).
- Cheney, J., Chiticariu, L., Tan, W.C. (2009). Provenance in Databases: Why, How, and Where. *Foundations and Trends in Databases*, 1(4). [DOI 10.1561/19000000006](https://doi.org/10.1561/19000000006).
- Gregor, S., Hevner, A.R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, 37(2), 337-355.
- Hevner, A.R., March, S.T., Park, J., Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, 28(1), 75-105.
- Larsen, K.R., Lukyanenko, R., Mueller, R.M., Storey, V.C., Parsons, J., VanderMeer, D., Hovorka, D.S. (2025). Validity in Design Science. *MIS Quarterly*, 49(4). [DOI 10.25300/MISQ/2024/18064](https://doi.org/10.25300/MISQ/2024/18064).
- Menotti, L., Marchesin, S., Giachelle, F., Silvello, G. (2025). Provenance-driven nanopublications: representing source lineage and trust networks for multi-source assertions. *International Journal on Digital Libraries*, 26, 24. [DOI 10.1007/s00799-025-00431-x](https://doi.org/10.1007/s00799-025-00431-x).

[Download PDF ↓](#)

