A Lightweight Index for Reproducible Data Pipelines

Department of Computer Science-style poster layout

E. Niemi J. Saarinen M. Okafor reproducibility@example.org

Research question

How can a project make its computational evidence easy to inspect without imposing a heavyweight platform on every contributor?

- one manifest for code, data, and environment;
- checksums for every published input;
- human-readable reports for failed checks.

Design principles

The index is deliberately boring: a small text manifest, deterministic validation, and a report that can be archived with the paper. This makes the useful path obvious when a build fails.

Workflow

Declare → Validate → Archive

manifest.toml report.html release.zip

Result at a glance

Across 42 public projects, the index found 96% of missing or stale inputs before publication. Median validation time was 14 seconds on a laptop, and every failure included a suggested repair command.

Measured outcomes

Measure	Before index	With index
Stale inputs caught pre-release	31%	**96%**
Median validation time	8 min	**14 s**
Projects with archived environment	24%	**88%**

What changed

The index does not replace peer review. It gives reviewers a compact map of what was run, which inputs were used, and which generated files can be recreated.

Limitations

Projects with proprietary data still need a safe redaction workflow. Network-dependent computations are recorded but not made deterministic by the index alone.

Next steps

1. add container provenance;
2. test multilingual manifests;
3. measure adoption after a six-month pilot.

Conclusion

Small, inspectable checks make reproducibility routine.

Original starter: replace this line with your project URL and QR code.