Paper 29 — Public Dataset Format

> Status: methodology paper, v0.7. Reading order: assumes 14-safety-and-blinding, 16-daq-dsp, 24-reproducible-build. Companion code: src/dataset.ltl — schema v1 bundle writer, signing chain, FAIR manifest generator.

1. Why this paper exists

At the end of the first-light campaign — regardless of the sign of the result — the full run dataset is released to a public archive (paper 28 §9.1). This paper defines exactly **what is in that archive**: the file format, the physical units, the signing chain that links every datum to the device key, the FAIR-data commitments, and the 50-year readability contract.

Without this paper, "data release" is ambiguous. A homodyne time-series is a sequence of numbers; without metadata it is uninterpretable. A PDF of our analysis results is not data. This paper specifies the format that makes the dataset interpretable by any competent physicist in 2026, 2036, or 2076.

2. Schema v1 — bundle layout

Every science run produces one `schema_v1` bundle. The bundle is a directory whose name is `e115d-run-<run_id>-schema_v1/`, containing:

```
e115d-run-<run_id>-schema_v1/
├── manifest.json          # signed manifest (§3)
├── run_header.json        # run-level metadata (§4)
├── raw/
│   ├── homodyne_I.f64le   # in-phase homodyne samples, §5
│   ├── homodyne_Q.f64le   # quadrature samples
│   ├── shot_noise_psd.f64le  # per-window shot noise, §5.3
│   └── timestamps.u64le   # sample timestamps, §5.4
├── ledger/
│   ├── run_ledger.cbor    # hash-chained ledger, paper 14
│   └── seal_certificate.json  # sealed run fields, paper 26 §2
├── calibration/
│   ├── nuisance_posteriors.json  # paper 22 §4 posteriors for this run
│   ├── cavity_characterisation.json
│   └── rf_chain_sweep.json
├── analysis/
│   ├── ln_b10_result.json  # primary result, §6
│   └── nested_sampling_chains.hdf5  # dynesty chains, §6
└── PROVENANCE.txt         # human-readable provenance note, §8
```

All files are generated by `dataset::write_bundle()` in src/dataset.ltl, which is called automatically at the end of each run's analysis phase. The bundle is written to the DAQ NAS and a SHA-256 is immediately appended to the run ledger.

3. The manifest and signing chain

`manifest.json` is the root of the signing chain. It is a JSON object:

```
{
  "schema_version": 1,
  "run_id": 11,
  "corpus_tag": "v0.7",
  "created_utc": "2026-04-28T14:22:00Z",
  "device_key_fingerprint": "A4:7C:...",
  "files": {
    "raw/homodyne_I.f64le":        "sha256:<hex>",
    "raw/homodyne_Q.f64le":        "sha256:<hex>",
    "raw/shot_noise_psd.f64le":    "sha256:<hex>",
    "raw/timestamps.u64le":        "sha256:<hex>",
    "ledger/run_ledger.cbor":      "sha256:<hex>",
    "ledger/seal_certificate.json": "sha256:<hex>",
    "calibration/...":             "sha256:<hex>",
    "analysis/ln_b10_result.json": "sha256:<hex>",
    "analysis/nested_sampling_chains.hdf5": "sha256:<hex>"
  },
  "manifest_signature": "<base64 Ed25519 signature over the above>"
}
```

The device key is an Ed25519 key generated during the Mk1 build and stored in the hardware security module (paper 24 §6). The signature covers the UTF-8 JSON of the manifest excluding the `manifest_signature` field itself (canonical form: sorted keys, no trailing whitespace). Any modification to any file in the bundle will cause the manifest to fail verification.

Verification command:

```
python tools/verify_bundle.py e115d-run-11-schema_v1/
```

Exit code 0 = valid; any other code = tampered or corrupted.

4. Run-header fields

`run_header.json` contains the complete set of run parameters as defined in paper 26 §2, plus:

Field	Type	Units	Description
`run_id`	u64	—	Monotone run counter
`phase`	string	—	Campaign phase (A–E)
`start_utc`	string	ISO-8601	Start of science window
`end_utc`	string	ISO-8601	End of science window
`sample_rate_hz`	f64	Hz	ADC sample rate after decimation
`n_samples`	u64	—	Total samples in homodyne_{I,Q}
`omega_m_hz`	f64	Hz	Mechanical resonance frequency
`f_sb_target_hz`	f64	Hz	Sideband target frequency
`kappa_hz`	f64	Hz	Cavity decay rate (full width)
`gamma_m_hz`	f64	Hz	Mechanical linewidth
`n_cav`	f64	photons	Mean intracavity photon number
`temperature_k`	f64	K	Cold-mass temperature at run start
`drive_alpha_sq`	f64	—	$
`noise_gate_pass`	bool	—	Whether §7 of paper 27 gates passed
`blinding_active`	bool	—	Whether analysis blind was active

All fields are required. A bundle with a missing required field is not schema-v1-conformant and must be rejected by any conformant reader.

5. Data files — units and encoding

5.1 `homodyne_I.f64le`, `homodyne_Q.f64le`

- Encoding: little-endian IEEE 754 double (8 bytes/sample)
- Units: **volts** at the homodyne photodetector output, calibrated against the shot-noise reference of §5.3
- Sample rate: as given in `run_header.json::sample_rate_hz`
- DC offset removed by the DSP pipeline (paper 16 §3) — each file has zero mean to within 10^{-6} V

5.2 Index convention

Sample k in `homodyne_I.f64le` (zero-indexed) corresponds to timestamp `timestamps.u64le[k]` (see §5.4) and to the same index k in `homodyne_Q.f64le`. The two files are strictly synchronous; they are not allowed to have different lengths.

5.3 `shot_noise_psd.f64le`

- One entry per DSP window (paper 16 §4)
- Units: V²/Hz, estimated from the out-of-band spectrum
- Used by `src/inference.ltl` as σ_k^2 in the likelihood (paper 22 §2)
- Length: `ceil(n_samples / window_size)` entries

5.4 `timestamps.u64le`

- Little-endian u64, one entry per sample
- Units: **nanoseconds since Unix epoch** (UTC, GPS-disciplined)
- Derived from the GPS-disciplined PPS reference (paper 08 §5)
- Absolute accuracy: < 50 ns; relative (between consecutive samples): < 200 ps

6. Analysis output

`analysis/ln_b10_result.json`:

```
{
  "run_id": 11,
  "ln_b10": 9.41,
  "ln_z1": -14502.3,
  "ln_z0": -14511.7,
  "verdict": "Hint",
  "n_live_points": 2000,
  "convergence_delta_ln_z": 0.08,
  "posterior_samples": 2000,
  "seal_hash": "<hex of SEALED_PRIORS_V0_5>",
  "analysis_binary_sha256": "<hex>"
}
```

`analysis/nested_sampling_chains.hdf5`:

- Standard `dynesty` HDF5 output format
- Includes the full posterior over $(a, \theta, \boldsymbol{\nu})$
- Reproducible from the sealed priors and the raw data by running `make reanalyse RUN=<run_id>` from the hand-off package root

7. FAIR commitments

This dataset is designed to be Findable, Accessible, Interoperable, and Reusable under the FAIR principles (Wilkinson et al., 2016):

Findable: - Each bundle is deposited to Zenodo with a persistent DOI. - The DOI is registered in the run ledger (paper 14) and in the hand-off package manifest (paper 28 §4.1). - All bundles from a single campaign are grouped under a single Zenodo community and a single collection DOI.

Accessible: - All bundles are deposited under CC0 (no rights reserved). - The Zenodo deposit is mirrored to OSF as a secondary archive. - The OT commits to maintaining the Zenodo deposit for a minimum of 20 years after campaign close.

Interoperable: - All numeric files use standard encodings (IEEE 754, little-endian; HDF5 for nested-sampling chains; JSON for metadata). - No proprietary formats. No binary-only tools are required to read the data. - Unit conventions follow SI throughout; all fields in the header include a units field.

Reusable: - `PROVENANCE.txt` (§8) gives human-readable context sufficient to interpret the data without any other document. - The schema version field (`schema_version: 1`) is part of the manifest; future schema versions will be backward-compatible with a documented mi-

gration path. - The sealed analysis code and sealed priors are included in every bundle, so any third party can reproduce the $\ln B_{10}$ result.

8. `PROVENANCE.txt`

Every bundle includes a plain-text file readable without any software:

```
element-115-drive run <run_id> — provenance note
=================================================
This bundle contains data from a cavity-optomechanical
propulsion-research experiment. The experiment tests a
speculative Mach-effect + photon–phonon propulsion model
described in docs/papers/07-mach-effect.md.

The primary result is analysis/ln_b10_result.json::ln_b10.
A value >= 15.4 constitutes a pre-registered discovery claim
(see docs/papers/22-bayesian-pipeline.md).

The raw data are homodyne quadratures of a cryogenic
photonic-crystal cavity. Units are volts at the detector;
the calibration chain is in calibration/cavity_characterisation.json.
Full documentation: https://github.com/bad-antics/element-115-drive
Corpus PDFs: dist/pdfs/ in the repository root.

SHA-256 of this bundle (manifest.json): <hex>
Signed by device key: <fingerprint>
Corpus tag: v0.7
```

`PROVENANCE.txt` is intentionally written at the level of a physicist who has never seen the project before. The 50-year readability commitment (§9) relies on this file being interpretable without any additional context.

9. Long-horizon readability

Data meant for replication are only useful if they remain readable. Commitments:

Horizon	Commitment
5 years	OT maintains a working `tools/verify_bundle.py` against current Python/OS
10 years	Zenodo deposit confirmed live; any URL changes updated in the run ledger
20 years	A static HTML rendering of each `run_header.json` and `ln_b10_result.json` is deposited alongside the bundle
50 years	`PROVENANCE.txt` and the raw data files (§5) are sufficient to reconstruct the primary result with any future tool chain; no OT-specific software is required

The 50-year commitment is achievable because: - IEEE 754 little-endian doubles are a 70-year-old standard with no sign of obsolescence. - HDF5 has ISO standardisation and multiple independent implementations. - JSON is an IETF standard (RFC 8259); Ed25519 is RFC 8032. - `PROVENANCE.txt` requires only a text editor.

10. Open questions

1. **Version 2 schema.** If the Mk2 architecture (paper 30) adds new data channels (e.g. gravimeter, secondary interferometer), a schema v2 will be required. Schema v2 must be a superset of v1; existing readers of v1 bundles must not break.
2. **Quantum-noise correlations.** For Mk2's back-action evasion mode (paper 30 §5), the I and Q channels are no longer independent and the shot-noise estimate in §5.3 becomes a matrix. The v2 schema should include a `shot_noise_covariance.f64le` field.
3. **Archive longevity beyond 20 years.** Zenodo's 20-year guarantee is institutional, not contractual. The OT will investigate Software Heritage and the Internet Archive as secondary mirrors before the campaign closes.

See 00-index for the corpus map.
