IEEE S&P 2026 · METACRISP WORKSHOP

From “Available” to “Reusable”

Measuring transparency and reproducibility of industrial control system datasets under security artifact evaluation policies

Mohammad Einaam Alim · **Tommy Morris**

Center for Cybersecurity Research & Education · University of Alabama in Huntsville

mohammad.alim@uah.edu · tommy.morris@uah.edu

Availability is a weak proxy for scientific reuse

AECs verify retrieval. They rarely verify that an evaluator can reconstruct what the data means.

WHAT AE POLICIES VERIFY TODAY

Artifact availability

USENIX Security ’25/’26: mandatory availability stage; optional functionality and reproducibility stage.

Olszewski et al. (2025): after AECs, artifact availability did not significantly change — but artifacts that go through AE function correctly at a higher rate.

Crowder et al. (2025): AECs encourage sharing but do not accurately assess shared datasets for reuse quality.

WHAT ICS DATASETS REQUIRE

Provenance and semantics

Cyber-physical traces are not intrinsically self-describing.

Scientific utility depends on:

- where signals were captured
- how variables map to physical state
- how perturbations were introduced
- how labels were derived

Without these, two groups citing the same dataset implement materially different splits, labels, and pipelines — and report incomparable results.

What an ICS / SCADA dataset actually contains

Two structural families dominate the corpus. Every release sits somewhere on this spectrum — and ships incomplete semantics.

FAMILY A

Process telemetry

Multivariate time series from sensors and actuators across a physical process.

timestamp	FIT-101	LIT-101	P-101	...	label
00:00:00	2.51	521.3	ON	...	normal
00:00:01	2.50	521.6	ON	...	normal
00:00:02	2.48	522.0	ON	...	attack

Examples: SWaT, WADI, HAI 1.0 / 20.07 / 21.03, BATADAL, EPIC.
Typical: 1 Hz sampling, 40–100 signals, days to weeks of operation, attack scenarios at known times.

FAMILY B

Network captures

Protocol-level traffic between PLCs, RTUs, HMIs, and historians on the OT segment.

time	src -> dst	proto	fn	register
0.0021	10.0.2 -> 10.0.7	Modbus	03	40001
0.0044	10.0.7 -> 10.0.2	Modbus	03	40001 OK
0.0067	10.0.9 -> 10.0.2	Modbus	06	40005 W

Examples: CSET Modbus 2016, CIC Modbus 2023, QUT DNP3, IEC 104.
Typical: PCAP + derived NetFlow / CSV; function codes, register reads / writes, packet inter-arrival times, optional labels.

WHAT THE FILES ALMOST NEVER SHIP

signal–semantics map · units, roles, subsystem mapping **clock semantics** · synchronization, drift, gap rules **event-aligned attack timeline** · on the dataset timebase **label transition rules** · onset / recovery / dwell

Three-tier outcome hierarchy for dataset–paper pairs

Grounded in badge semantics already used by security AECs; the unit of analysis is one dataset paired with one paper.

01

Available

Evaluator retrieves the dataset via the stated process and unambiguously identifies what was retrieved.

MINIMUM EVIDENCE

Stable identifier (DOI / tag / checksum); access record.

02

Functional for Reuse

Evaluator parses the dataset and reconstructs schema, label, split, and preprocessing semantics sufficient to run at least one baseline or derive a claim-relevant invariant.

MINIMUM EVIDENCE

Signal-semantics map; explicit label rules; split protocol; preprocessing definitions.

03

Reproduced

At least one dataset-dependent claim from the paired paper is reproduced within a pre-declared tolerance.

MINIMUM EVIDENCE

Pre-declared targets and tolerances; rerun logs; stochasticity controls (seeds, trials).

Outcomes are ordered by strength. Reproducibility is not a property of files; it is a property of the system of data, pipeline, documentation, and stated claims.

TS-ICS — six provenance dimensions, each scored 0 / 1 / 2

Criterion-based, adapted from transparency metascience in usable privacy and security (Klemmer et al., 2025).

SYS System Provenance

Topology, components, signal-semantics map: variable meaning, units, role, subsystem.

CAP Capture Provenance

Capture points, sampling rates, timestamp semantics, clock and synchronization assumptions.

ATT Perturbation / Attack

Mechanism, affected components, event timelines aligned to the dataset timebase.

LAB Label Semantics

Granularity, class definitions, transition rules for onset and recovery.

PRE Preprocessing & Features

Windowing, normalization fitting partition, split protocol, leakage controls, features.

REL Release & Archival

License, versioning, stable identifiers, integrity checks, pinned environments.

SCORING

0 absent • **1** partial — evaluator must infer key elements • **2** complete — unambiguous reconstruction from public evidence

Total TS-ICS score ∈ [0, 12]. Rubric supports inter-rater reliability checks within AECs.

A nine-stage reproducibility workflow with structured logging

Stages 0 – 8. Every step produces a persistent record an independent evaluator can audit.

PHASE 1 · Context & Semantics

0 Freeze evaluation context
OS image, pinned deps, hardware, nondeterminism notes.

1 Acquisition & integrity
Access record, DOI/checksum, time-to-access.

2 Semantic reconstruction
Schema + signal-semantics map; assumption ledger.

Yields "Available"

PHASE 2 · Provenance & Pipeline

3 Capture provenance
Capture points, sampling, timestamp semantics.

4 Perturbation alignment
Event-aligned attack table on dataset timebase.

5 Label semantics
Taxonomy, inclusion rules, transition rules.

6 Preprocessing pipeline
Windowing, normalization, splits, leakage controls.

Yields "Functional for Reuse"

PHASE 3 · Targets & Outcomes

7 Declare reproduction targets
Pre-declared tolerances; seed and trial controls.

8 Outcome assignment
TS-ICS vector, blocker codes, auditable log.

Yields "Reproduced"

15 highly-cited ICS dataset–paper pairs, 2011 – 2025

Diverse panel that exercises the rubric across domains, access tiers, and data modalities. Not exhaustive.

15
dataset–paper pairs
public evidence only

4
process domains
water · power · IIoT · SCADA/protocols

6
TS-ICS dimensions
SYS · CAP · ATT · LAB · PRE · REL

3
outcome tiers
Available · Functional · Reproduced

INCLUSION CRITERIA

- At least one empirical claim in the paired paper depends on the dataset.
- Dataset relevant to ICS or SCADA.
- Artifacts accessible under a stated policy — open or request-gated.
- Stably identifiable by release tag, DOI, or checksum (release-quality flag otherwise).

Sources: Google Scholar, IEEE Xplore, USENIX, ACM DL. Query families include ICS, SCADA, Modbus, DNP3, IEC 60870-5-104, testbed, benchmark.

TS-ICS scores across the corpus

Scores from publicly available documentation and artifacts. 0 absent · 1 partial · 2 complete.

Dataset	Domain	SYS	CAP	ATT	LAB	PRE	REL	Total	Outcome
SWaT	Water	2	2	1	1	1	1	8	Functional
WADI	Water	2	2	1	1	1	1	8	Available
BATADAL	Water	1	1	2	1	1	2	8	Functional
EPIC	Power	2	1	1	0	0	1	5	Available
HAI 1.0	Power	2	2	2	1	1	2	10	Functional
HAI 20.07	Power	2	2	2	2	1	2	11	Functional
HAI 21.03	Power	2	2	2	2	2	2	12	Reproduced
WUSTL-IIoT-2018	IIoT	1	1	1	1	1	1	6	Available
CSET Modbus 2016	SCADA	1	1	1	1	1	1	6	Functional
CIC Modbus 2023	Modbus	1	2	2	2	1	2	10	Functional
Cyber4OT	OT net	1	1	2	1	1	1	7	Functional
ICS Coll. (Morris)	SCADA	1	1	1	1	1	1	6	Functional
QUT DNP3	DNP3	1	1	1	1	1	1	6	Functional
IEC 60870-5-104	IEC 104	1	2	2	2	1	2	10	Functional
SCADA Testbed Coll.	SCADA	1	2	2	2	1	1	9	Functional

Across the corpus: 1 Reproduced · 11 Functional for Reuse · 3 Available. Preprocessing is the weakest category on average.

Reuse collapses at preprocessing and label semantics

Datasets are accessible but rarely auditable as scientific objects — even when code runs.

PRE — the preprocessing blind spot

Weakest category on average

- Normalization fitting partition almost never declared.
- Windowing strides and overlap rules under-specified.
- Leakage controls between train/val/test absent.
- Hidden filtering rules invalidate downstream metrics.

LAB — ambiguous label semantics

Frequent partial scores

- Onset and recovery transition rules under-defined.
- Inclusion / exclusion criteria implicit.
- Label files often inconsistent with event logs.
- Class taxonomy not regenerable from public evidence.

THE PARADOX

Authors specify deep learning architectures to four decimal places and treat dataset construction — splits, normalization, labels — as an afterthought. A dataset can be public, accessible, and still fail functional reuse.

From availability toward reusable, comparable science

For AECs

Move from availability verification to semantic validation. Require a signal-semantics map and an explicit split schema as first-class artifacts.

For Authors

Publish deterministic data-generation and preprocessing scripts alongside models. Treat the pipeline — not just the CSV — as the artifact under review.

For Venues

Recognize that a shared file without physical context is an incomplete artifact. Set expectations that match the cyber-physical nature of the data.

OPEN CHALLENGE FOR ARTIFACT EVALUATION COMMITTEES

If an artifact's physical semantics cannot be reconstructed without insider knowledge, should the paper be eligible for a Functional or Reproduced badge — even if the code runs perfectly?