Towards Trust Engineering in Open Data Systems: A Layered Conceptual Framework Integrating Quality Assurance and Governance Perspectives

Luciano Santos Pinheiro, Cristiane de Holanda de Barros e Silva, Vitor Barros Aquino,
Thays Maria da Conceicao Silva Carvalho, Cristiano Vale do Rego Barros Filho, Washington Henrique Carvalho Almeida

Presenter: Luciano Santos Pinheiro
Affiliation: Cesar School, Recife, Brazil
Email: lsp3@cesar.school

About the Presenter

Luciano Santos Pinheiro

Manager, SIGEC/IBAMA | MSc Student, Cesar School

EDUCATION

B.Sc. in Psychology, UNICEUMA University, Sao Luis, Maranhao, Brazil (2003)
Currently pursuing M.Sc. in Software Engineering, Cesar School, Recife, Pernambuco, Brazil

PROFESSIONAL

Since 2006: Environmental analyst at IBAMA (Brazilian Institute of Environment and Renewable Natural Resources)
Current role: Manager, Innovation and Knowledge Management Service (SIGEC)

RESEARCH INTERESTS

Innovation | Artificial intelligence | Local LLMs | Big data | Data quality | Open data | Trust engineering | Environmental issues

The Trust Deficit in Open Data Systems

THE PROBLEM

- Open data systems are critical infrastructure for democratic governance and civic participation
- Persistent **trust deficits** undermine their potential value

ROOT CAUSES

- **Silent quality regressions** — undetected schema drift
- **Inadequate provenance tracking** — no lineage records
- **Fragmented approaches** — quality OR governance, never both

"A unified conceptual model integrating quality assurance mechanisms with governance principles remains absent."

EXISTING FRAMEWORKS COVER ONLY PART OF THE PROBLEM

Q ISO 25012, Wang & Strong — quality taxonomies only

G Data trust models — governance structures only

! **Gap:** No unified Q + G integration model

FAIR compliance ≠ Data quality

(Nicholson et al., 2025)

Wirtz et al., 2019 | Attard et al., 2015 | Nicholson et al., 2025

Our Contribution

1. **Five-Layer Conceptual Framework**
   Integrating quality assurance mechanisms with governance principles hierarchically

2. **Terminological Clarification**
   Distinguishing Verification ("built right?") from Validation ("built the right thing?") in data QA

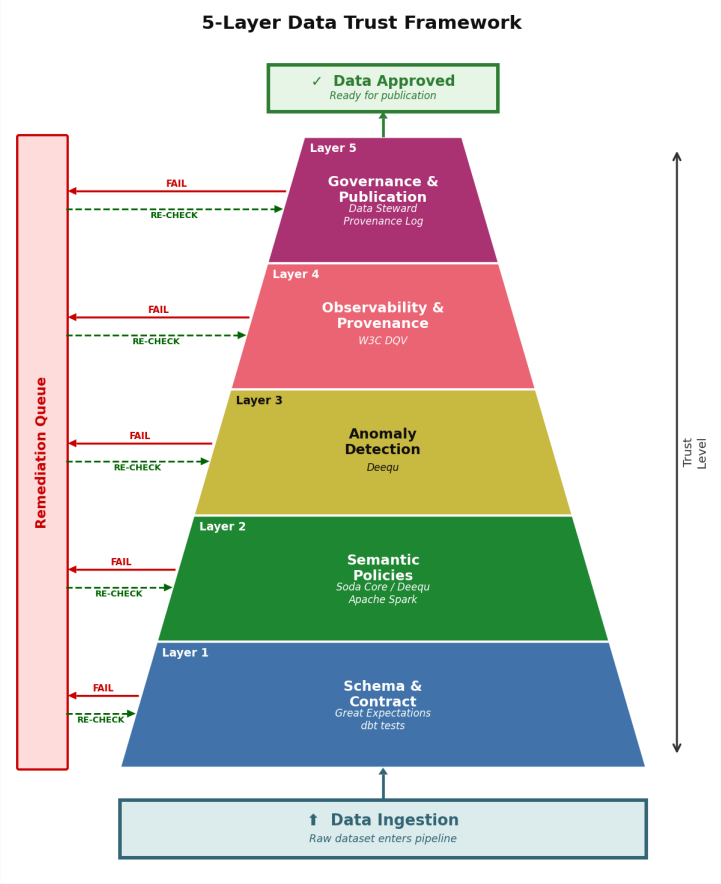
3. **Comparative Analysis**
   Positioned against ISO 25012, FAIR, W3C DQV, and Apache Deequ

4. **Illustrative Application**
   Applied to IBAMA pesticide sales dataset — real open government data

5. **Six Testable Propositions**
   Linking framework adoption to trust outcomes for future empirical investigation

Index Terms:
data quality | trust engineering | open data | data governance | FAIR principles

Five-Layer Trust Engineering Framework



L5 — **Governance**
Publication gate, steward roles, policy enforcement

L4 — **Observability & Provenance**
W3C PROV-DM, lineage, transformation history

L3 — **Anomaly Detection**
Statistical monitoring, drift detection, Z-score thresholds

L2 — **Semantic Policies**
Business rules, domain-specific constraints, Soda Core

L1 — **Structural Contracts**
Schema, types, required fields — Great Expectations

Pinheiro et al., 2026 | Beck, 2003 | Meyer, 1992

Validation Layers: From Structure to Behaviour

L1 Structural Contracts

DESIGN-BY-CONTRACT PRINCIPLE

- Machine-readable schema specifications
- Prevents schema drift and type errors
- Enforces nullability and referential integrity

Tools: Great Expectations, Soda Core, dbt

L2 Semantic Policies

DOMAIN-SPECIFIC VALIDATION

- Business rules beyond syntactic correctness
- Catches logical errors invisible to structural checks
- Requires domain expert involvement

Tools: Soda Core (SodaCL), custom rule engines

L3 Anomaly Detection

BEHAVIOURAL MONITORING

- Z-score thresholds over rolling 12-period baseline
- Page-Hinkley test for concept drift detection
- Two-stage: automated flag + steward review

Tools: Apache Deequ, MOA framework

Beck, 2003 | Meyer, 1992 | Majors et al., 2022

Observability, Provenance & Governance

L4 Observability & Provenance

- W3C PROV-DM: directed acyclic graph of entities, activities, agents
- Mandatory provenance fields: source_uri, ingest_timestamp, checksum_sha256, transformation steps
- W3C DQV: quality metadata published as linked data
- Enables transparency, auditability, and informed trust decisions

Addresses: credibility, traceability, understandability dimensions

L5 Governance Layer

- Publication gate: authority to halt publication pending remediation
- Steward roles with defined accountability structures
- Translates quality signals into organizational decisions
- Bridges technical QA and stakeholder accountability

Trust = Technical mechanisms (L1-L4) + Organizational governance (L5)

Cheney et al., 2009 | Groth & Moreau, 2013 | Majors et al., 2022

Case Study: IBAMA Pesticide Sales Dataset

Dataset: IBAMA Pesticide Sales 2007-2023

33 columns | 27 state-level sales columns (tonnes) | Ingrediente_ativo, Ano, Semestre

L1 — Structural Contracts

All 33-column structural checks PASSED — encoding, schema, integer types confirmed

L2 — Semantic Policies

10,959 records with negative sales values flagged — escalated to steward (Layer 5)

L4 — Provenance Gap

Undocumented schema evolution: 2 semesters/yr (2007-2021) to 1 (2022+) — lineage gap identified

L3 — Anomaly Detection Finding

Glyphosate sales decline 2022 to 2023

-15.6% decline

Classified as natural behavior (regulatory changes)

L5 — 3 Governance Decisions Required:

(i) Classify negative-value records
(ii) Document 2022 semester-structure change
(iii) Establish provenance publication policy

Positioning Against Existing Frameworks

Positioning the framework against established models across three dimensions

Dimension	Proposed	ISO 25012	FAIR	W3C DQV	Deequ
Quality Dimensions	**Full**	**Full**	Partial	Partial	Partial
Governance Integration	**Full**	None	Partial	None	None
Operational Guidance	**Full**	None	None	Partial	**Full**
Continuous Validation	**Full**	None	None	None	**Full**
Trust Mechanisms	**Full**	None	Partial	Partial	None

ISO/IEC, 2008 | Vetro et al., 2016 | Wilkinson et al., 2016

Six Testable Propositions

P1 **Trust & Transparency**

Observability mechanisms (L4) will exhibit higher stakeholder trust, mediated by perceived transparency

P2 **Quality & Validation Layers**

Multiple validation layers will demonstrate fewer quality defects, with diminishing returns beyond three layers

P3 **Governance & Quality Signals**

Integrating quality signals into publication decisions (L5) will yield more consistent quality over time

P4 **Anomaly Detection & Timeliness**

Continuous anomaly detection (L3) will reduce mean time to quality degradation detection by at least 50%

P5 **Semantic Policies & Domain Expertise**

Effectiveness of semantic policy layers (L2) is positively moderated by domain expert involvement

P6 **Framework Adoption & Maturity**

Higher governance maturity predicts hierarchical layer adoption (L1-L5); lower maturity leads to opportunistic adoption

Pinheiro et al., 2026 | Milne et al., 2022 | Radosevic et al., 2023

Theoretical & Practical Implications

THEORETICAL IMPLICATIONS

Bridges Quality Theory + Trust Theory

Trust as a systemic property emerging from technical mechanisms (L1-L4) and organizational processes (L5) — maps to Mayer et al.'s competence, integrity, and benevolence dimensions

Addresses Gaps in Existing Integrations

DAMA-DMBOK: process maturity, not validation architecture. DataOps: pipeline velocity, not governance. Data mesh: ownership, not inter-domain quality contracts

Operationalizes Trust Principles

Links quality assurance mechanisms to transparency, accountability, and informed trust decisions through concrete technical layers

PRACTICAL IMPLICATIONS

Hierarchical Adoption Roadmap

Start with L1 (structural contracts), then add L2, L3, L4 incrementally — gain value at each stage before investing in higher layers

Technology Mapping

L1: Great Expectations / Soda Core | L2: Rule engines | L3: Apache Deequ / MOA | L4: W3C PROV-DM | L5: Governance policies

Staged Approach for Resource-Constrained Orgs

Enables incremental adoption aligned with organizational maturity — not all layers required simultaneously

Pinheiro et al., 2026 | Wang & Strong, 1996 | Milne et al., 2022

Limitations & Future Research Directions

LIMITATIONS

- Conceptual framework — requires empirical validation through case studies and controlled experiments
- May underweight social and organizational factors in trust-building (socio-technical gap)
- Limited applicability to unstructured data (text, images, video)
- Does not address adversarial scenarios or data integrity attacks
- Assumes sufficient technical capacity — may be infeasible for resource-constrained organizations

FUTURE RESEARCH DIRECTIONS

1. Empirical validation through longitudinal case studies across diverse organizational contexts — test all six propositions
2. Open-source reference implementation: Great Expectations + Soda Core + dbt + W3C PROV-DM against IBAMA dataset
3. Extension to streaming data, federated systems, and AI training datasets (concept drift, distributed validation)
4. Investigation of organizational factors: culture, governance maturity, DataOps integration, data mesh architectures

Conclusion

- ✓ Proposed a **five-layer conceptual framework** for trust engineering in open data systems, integrating quality assurance mechanisms with governance principles hierarchically
- ✓ Identified **three critical gaps** in existing frameworks: fragmentation between quality and governance, lack of continuous validation guidance, insufficient attention to temporal quality dimensions
- ✓ Positioned the framework within ISO 25012, FAIR, W3C DQV and Deequ landscapes — **only framework offering full coverage** across all five evaluated dimensions
- ✓ Articulated **six testable propositions** providing a concrete research agenda for empirical validation across diverse organizational contexts

Contributes a conceptual foundation for engineering trustworthy open data systems that balances transparency, risk management, and stakeholder accountability

References

Pinheiro, L. S., Silva, C. H. B., Aquino, V. B., Carvalho, T. M. C. S., Barros Filho, C. V. R., & Almeida, W. H. C. (2026). Towards trust engineering in open data systems: A layered conceptual framework integrating quality assurance and governance perspectives. *Proceedings of IARIA SoftEng 2026 Congress*.

ISO/IEC. (2008). *ISO/IEC 25012:2008 — Software Engineering — Data Quality Model*. International Organization for Standardization.

Milne, R., Morley, J., & Howard, H. S. (2022). Data trusts and the governance of health data. *Nature Medicine*, *28*, 2218-2220.

Vetro, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open data quality measurement framework. *Government Information Quarterly*, *33*(2), 325-337.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5-33.

Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18