# 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 Conceição Silva Carvalho,
Cristiano Vale do Rego Barros Filho, Washington Henrique Carvalho Almeida
Cesar School
Recife, Brazil
email: {lsp3, chbs, vba, tmcsc, cvrbf, whca}@cesar.school

***Abstract*—Open data systems face persistent trust deficits due to silent quality regressions, schema drift, and inadequate provenance tracking. While existing frameworks address either data quality measurement or governance structures in isolation, a unified conceptual model integrating quality assurance mechanisms with governance principles remains absent. This paper proposes a five-layer conceptual framework for trust engineering in open data systems, synthesizing insights from data quality theory, data trust models, and software engineering validation practices. The framework organizes quality assurance mechanisms into a hierarchical pyramid—from structural contracts to semantic policy checks, anomaly monitoring, and observability—with each layer addressing distinct quality dimensions while collectively building trust through transparency and accountability. We position this framework within existing theoretical landscapes, including Findability, Accessibility, Interoperability, and Reusability (FAIR) principles, data trust governance, and International Organization for Standardization (ISO) quality standards, demonstrating how it extends current models by explicitly linking quality dimensions to executable validation mechanisms and publication governance decisions. Through comparative analysis of existing frameworks, we identify gaps in operationalizing trust through continuous validation and propose testable propositions for future empirical investigation. This work contributes a conceptual foundation for engineering trustworthy open data systems that balances transparency, risk management, and stakeholder accountability.**

***Index Terms*—data quality; trust engineering; open data; data governance; FAIR principles.**

## I. Introduction

We present a conceptual framework for trust engineering in open data systems. Throughout this paper, *trust engineering* refers to the systematic application of engineering practices to produce verifiable trust properties in data systems; *trust-building* refers to the organizational and social processes through which stakeholders develop confidence in those systems; and *trustworthy systems* refers to the resulting artefacts whose properties have been engineered and verified. These three concepts are complementary and collectively necessary for open data trust.

### A. Motivation and Problem Statement

Open data systems have become critical infrastructure for democratic governance, policy-making, and civic participation. However, persistent trust deficits undermine their potential value. Silent quality regressions—including schema drift, unstable identifiers, distribution shifts, and missing provenance—erode user confidence and limit data reuse [1], [2]. Empirical studies in software engineering corroborate this: Wu et al. demonstrate that silent label-quality errors in software datasets propagate undetected through automated pipelines, causing systematic downstream failures [41]. Unlike traditional software systems, where test-first methodologies have proven effective for quality assurance, open data ecosystems lack comparable conceptual frameworks that integrate continuous validation with governance structures.

Open data portals maintained by public agencies face additional challenges: heterogeneous consumer populations with varying technical expertise, absence of service-level agreements, and regulatory transparency requirements that demand auditability of quality decisions. The concept of *data trust*—defined here as the justified belief that a dataset accurately represents the phenomenon it purports to describe, and that its provenance and transformation history are transparent and auditable—is an engineered property produced by verifiable processes, not assumed by goodwill.

Existing theoretical work addresses data quality measurement [3], [4] and governance models [5]–[7], [29], [30] in isolation, but fails to provide an integrated conceptual foundation for engineering trust through systematic quality assurance. Quality frameworks such as ISO 25012 [3], the W3C Data Quality Vocabulary (DQV) [37], and the FAIR principles [16] enumerate desirable properties but stop short of prescribing the engineering mechanisms through which those properties are achieved and maintained over time. Data trust models emphasize participatory governance and stakeholder engagement [5], [8], while quality frameworks focus on dimensional assessment [3], [4], [9]. This fragmentation leaves practitioners without clear guidance on how quality assurance mechanisms



 ISBN: 978-1-68558-398-9 

relate to trust-building practices and governance decisions.

### B. Research Gap and Contribution

This paper addresses the gap between quality measurement and trust governance by proposing a five-layer conceptual framework that integrates quality assurance mechanisms with governance principles. Our framework synthesizes insights from software engineering validation practices [10], data quality theory [3], [9], and data trust models [5], [6] to provide a unified conceptual foundation for engineering trustworthy open data systems.

We contribute: (1) a five-layer conceptual framework integrating quality assurance with governance, with each layer mapped to specific quality dimensions, implementation technologies [22], [38], [40] and governance responsibilities; (2) a terminological clarification distinguishing verification ("Did we build the system right?") from validation ("Did we build the right system?") [20], [21] applied to data quality assurance; (3) a comparative analysis positioning the framework against ISO 25012, FAIR, W3C DQV, and Apache Deequ; (4) an illustrative application to the Brazilian Institute of Environment and Renewable Natural Resources (IBAMA) pesticide sales dataset [24]; and (5) six testable propositions linking framework adoption to trust outcomes.

The remainder of this paper is structured as follows. Section II reviews background and theoretical foundations. Section III describes the five-layer framework. Section IV presents the IBAMA illustrative application. Section V presents comparative analysis. Section VI presents six research propositions for empirical validation. Section VII discusses implications and limitations. Section VIII concludes.

## II. Background and Theoretical Foundations

We organize related work along four dimensions: (i) data quality measurement frameworks and standards; (ii) data trust and governance models; (iii) software engineering validation practices; and (iv) FAIR principles and data stewardship. We then identify the gap that our framework addresses.

### A. Data Quality Dimensions and Standards

Data quality research has established multidimensional frameworks for assessing fitness for use. Wang and Strong's seminal work identified fifteen quality dimensions organized into intrinsic, contextual, representational, and accessibility categories [9]. ISO/IEC 25012 standardized quality characteristics including accuracy, completeness, consistency, and credibility [3]. These frameworks provide taxonomies for quality assessment but lack operational guidance on implementing continuous validation mechanisms.

At the implementation level, several open-source tools operationalize quality measurement. Great Expectations [38] introduces *expectations*—declarative assertions about data properties evaluated at runtime. Soda Core [39] adopts a domain-specific language (SodaCL) for defining checks embeddable in orchestration pipelines. The dbt testing framework [40] integrates schema and referential integrity tests directly into transformation workflows. Apache Deequ provides a Scala/Spark-based library for automated quality monitoring at scale. Recent work extends dimensional models to open data contexts. Vetrò et al. proposed quality metrics tailored to open government data, emphasizing timeliness, accuracy, and accessibility [4]. Gong et al. confirm that completeness, consistency, and timeliness remain the most operationally critical dimensions while identifying the absence of integrated governance mechanisms as a persistent gap [42].

### B. Data Trust and Governance Models

Data trust frameworks emphasize governance structures that enable stakeholder participation and accountability. Milne and Brayne's data trust model proposes independent stewardship, participatory governance, and transparent decision-making as trust-building mechanisms [5]. Radosevic et al. extend this model to spatial data infrastructures [6]. Artyushina's civic data trust framework highlights transparency, accountability, and community participation [8]. The UK Food Standards Agency's food data trust initiative further demonstrates how sector-specific governance structures can operationalize data stewardship principles in practice [32].

Data mesh architectures [35] distribute governance responsibility to domain teams, introducing data products with embedded quality contracts. DataOps [31] applies continuous integration principles to data pipelines. While these approaches advance practice, they do not provide a unified theoretical model mapping specific quality dimensions to specific governance decisions. The present framework addresses this gap by making the governance decision the mandatory final step in the publication pipeline.

### C. Software Engineering Validation Practices

The distinction between verification and validation, introduced by Boehm [20] and formalized in IEEE Std 1012 [21], is central to the framework's design. Verification asks: "Are we building the product right?"—it checks conformance to a specification (e.g., does the dataset schema match the published contract?). Validation asks: "Are we building the right product?"—it checks fitness for the intended use (e.g., do the pesticide sales figures accurately reflect market reality?). In the context of data quality assurance, Layers 1–3 of our framework are primarily verification activities; Layer 4 supports both; Layer 5 is a validation activity in which human stewards exercise judgment about fitness for publication.

Software engineering has developed mature practices for continuous validation. Test-driven development establishes executable specifications that prevent regressions [10]. Design by contract formalizes preconditions, postconditions, and invariants as enforceable constraints [11]. Observability engineering provides runtime visibility into system behavior through structured logging, metrics, and tracing [12].

These practices have proven effective for maintaining software quality but have not been systematically adapted to open

data contexts. Our framework bridges this gap by translating software validation concepts to data quality assurance.

### D. FAIR Principles and Data Stewardship

The FAIR principles—Findability, Accessibility, Interoperability, and Reusability—provide foundational guidelines for scientific data management [16]. Nicholson et al. demonstrate that FAIR compliance does not imply quality: a dataset can be fully FAIR-compliant yet contain semantic errors, distributional anomalies, or outdated provenance records [17]. This observation motivates the explicit inclusion of a quality assurance layer beneath the governance layer: FAIR addresses discoverability and accessibility, whereas our framework addresses fitness for use and the engineering processes that sustain it.

Quality frameworks such as ISO 25012 [3], the W3C Data Quality Vocabulary (DQV) [37], and the FAIR principles [16] each address a subset of the quality-governance space but remain high-level guidelines requiring operationalization through specific technical mechanisms.

## III. Five-Layer Conceptual Framework

Our framework organizes quality assurance mechanisms into five hierarchical layers, each addressing distinct quality dimensions while collectively building trust through continuous validation and transparent reporting. The layers progress from low-level syntactic contracts to high-level semantic and organizational controls, mirroring the Open Systems Interconnection (OSI) network model's principle of layered abstraction. Each layer addresses a distinct class of quality failures; higher layers assume the guarantees provided by lower layers. Conflicts are resolved by the governance layer (Layer 5), which has authority to halt publication pending remediation. Alternative decompositions—by quality dimension, stakeholder role, or data lifecycle stage—were considered but rejected in favour of the hierarchical technical-to-governance ordering, which reflects natural implementation dependencies and supports incremental adoption: organisations can implement lower layers first and gain immediate value before adding higher layers.

Figure 1 illustrates both the framework architecture and the validation pipeline: the pyramid layers represent the five trust levels (L1–L5), and the Remediation Queue panel on the left shows how failed checks at each layer are routed for corrective action before re-ingestion.

### A. Layer 1: Structural Contracts

Structural contracts establish foundational guarantees about data schema, types, and required fields. Drawing from design-by-contract principles [11], this layer defines machine-readable specifications that prevent schema drift and ensure structural consistency. Structural contracts address the accuracy, completeness, and consistency dimensions by enforcing type constraints, nullability rules, and referential integrity.

Implementation technologies include: Great Expectations [38] `ExpectationSuite` objects serialised as JSON (versionable alongside the dataset); OpenAPI [25] schemas for

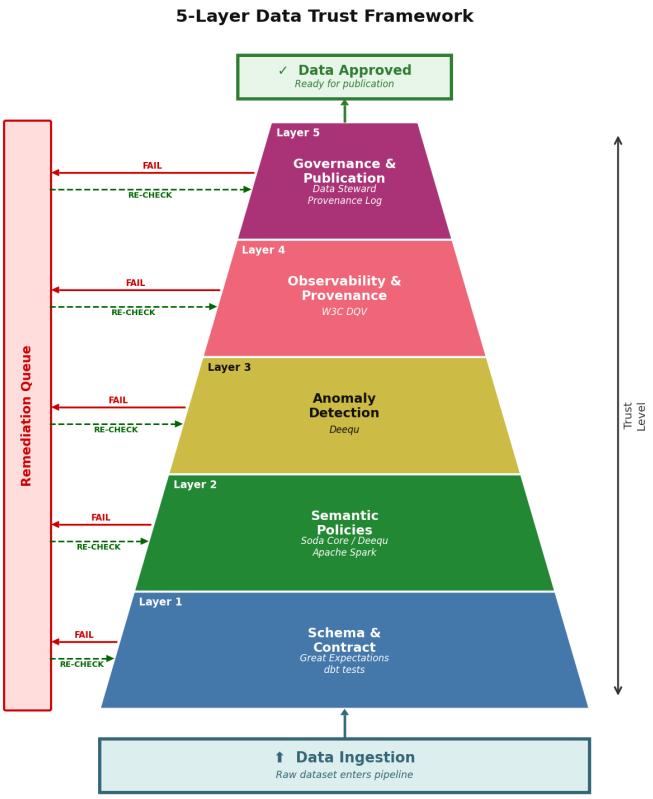


Fig. 1. Five-layer trust engineering framework with integrated remediation pipeline. Trust accumulates from structural contracts (Layer 1) to governance decisions (Layer 5); failed checks at each layer route to the Remediation Queue (left panel) for automated or manual correction before re-ingestion. Implementation tools are shown per layer.

API-delivered data; and dbt [40] schema YAML files defining column-level constraints evaluated on every pipeline run. These contracts serve as executable documentation, enabling automated validation at ingestion and publication boundaries.

### B. Layer 2: Semantic Policies

Semantic policies enforce domain-specific business rules and logical constraints that transcend structural validation. This layer addresses semantic accuracy and logical consistency by validating relationships between fields, enforcing domain constraints, and detecting logical inconsistencies. Semantic policies translate domain knowledge into executable rules that prevent semantically invalid data from propagating through systems.

Examples include range constraints (e.g., pesticide sales volumes must be non-negative for final annual records), cross-field validations (e.g., end date must follow start date), entity resolution rules [22], and controlled-vocabulary conformance. Soda Core [39] SodaCL checks express semantic rules in a human-readable domain-specific language evaluated at runtime; Apache Deequ `Check` objects support constraint verification at scale on Apache Spark. Semantic policies require domain expertise to define but provide critical protection against logically inconsistent data that passes structural validation.

This layer bridges the gap between syntactic correctness and semantic validity.

### C. Layer 3: Anomaly Detection

Anomaly detection monitors distributional properties and temporal patterns to identify unexpected changes in data characteristics. This layer addresses the timeliness, consistency, and credibility dimensions by detecting distribution shifts, outliers, and temporal anomalies that may indicate quality degradation or upstream process changes [26], [27].

The framework employs a two-stage approach: (i) automated statistical monitoring using control charts or Z-score thresholds to flag candidate anomalies, and (ii) steward review to classify flagged observations as natural behavior (document and accept), data error (remediate), or genuine anomaly (quarantine and investigate). This design prevents the framework from suppressing real signals while ensuring genuine errors are not published. Apache Deequ provides automated anomaly detection; the MOA framework [28] supports concept-drift detection for streaming data. Thresholds are determined per column using a rolling baseline window (default: 12 periods) and flagging observations that deviate by more than $k\sigma$ from the baseline mean, where $k$ is set by the data steward based on acceptable false-positive rates (recommended $k = 3$ for initial deployment). Concept drift—a sustained shift in the underlying data distribution rather than an isolated anomaly—is distinguished from point errors by applying the Page-Hinkley test [28] over the same rolling window; a confirmed drift triggers a schema evolution review rather than a remediation action. Anomaly detection complements rule-based validation by identifying quality issues that cannot be anticipated through explicit constraints. This layer provides early warning of quality degradation before downstream impacts occur.

### D. Layer 4: Observability and Provenance

Observability mechanisms provide transparency into data lineage, transformation history, and quality metrics. Drawing from observability engineering [12] and provenance research [13]–[15], this layer addresses the credibility, traceability, and understandability dimensions by documenting data origins, transformations, and quality assessments.

Provenance tracking captures who produced data, when, through what processes, and with what quality characteristics. Provenance records include: source system identification, ingestion timestamps, transformation steps with input/output checksums, quality check results from Layers 1–3, and steward decision records from Layer 5. The W3C PROV-DM standard [15] provides vocabulary for representing provenance as a directed acyclic graph of entities, activities, and agents. The W3C DQV [37] provides complementary vocabulary for publishing quality metadata as linked data. A minimal provenance record for each dataset version includes the following mandatory fields: `source_uri` (origin endpoint), `ingest_timestamp` (ISO 8601), `checksum_sha256` (structural fingerprint), `l1_pass`/`l2_pass`/`l3_pass` (boolean layer results), `anomaly_flags` (flagged column–period pairs), and `steward_decision` (approve, quarantine, or remediate). The observability dashboard exposes these fields as a time-series quality log, enabling consumers to inspect quality history and compare metrics across releases. Observability transforms opaque data pipelines into transparent, auditable systems.

### E. Layer 5: Governance and Publication Decisions

The governance layer integrates quality signals from lower layers into publication and access control decisions. This layer addresses accountability, compliance, and risk management by establishing thresholds for publication, defining stakeholder roles, and implementing feedback mechanisms. Governance policies translate quality assessments into actionable decisions about data release, access restrictions, and quality warnings.

The governance model specifies: (i) a *data steward* role responsible for reviewing quality check results and making publication decisions; (ii) a *data owner* role responsible for defining quality policies and accepting residual risk; (iii) a *data consumer* role with the right to access quality metadata and provenance records; and (iv) a change management process for schema and policy updates. Publication decisions are recorded in the provenance log (Layer 4), creating an auditable governance trail. The primary advantage of integrating quality measurement (Layers 1–4) with governance (Layer 5) is the elimination of the accountability gap: no accountable actor can bypass the quality evidence review before publication. Publication eligibility is determined by a composite Global Quality Score $Q_s = w_1 L_1 + w_2 L_2 + w_3 L_3 + w_4 L_4$, where $L_i \in [0, 1]$ is the pass rate for layer $i$ and weights $w_i$ sum to 1. Default weights ($w_1 = 0.3, w_2 = 0.3, w_3 = 0.2, w_4 = 0.2$) reflect the foundational importance of structural and semantic layers; the data owner has authority to adjust weights to reflect domain-specific risk tolerance. A dataset is eligible for publication if $Q_s \geq \theta$, where the publication threshold $\theta$ (default $\theta = 0.85$) is set by the data owner and reviewed annually or following any governance incident.

This layer operationalizes data trust principles [5], [6] by connecting technical quality mechanisms to organizational accountability structures. Governance frameworks define who can publish data, under what quality conditions, with what transparency requirements, and through what stakeholder engagement processes. This layer closes the loop between quality assurance and trust governance.

## IV. Illustrative Application: IBAMA Pesticide Sales Dataset

To demonstrate practical applicability, we apply the framework to the IBAMA pesticide sales dataset [24], a publicly available open data artefact published by the Brazilian federal environmental agency. The dataset contains 124,245 records covering 584 active ingredients across 19 years (2007–2026), distributed across 27 Brazilian states, encoded in UTF-8, semicolon-delimited, with 33 columns including

`Ingrediente_ativo`, `Ano`, `Semestre`, and 27 state-level sales columns measured in tonnes of commercial product.

**Layer 1 (Structural Contracts):** A Great Expectations `ExpectationSuite` was defined specifying 33 columns in fixed order, UTF-8 encoding, semicolon delimiter, and integer type for `Ano` (range 2007–2026). All structural checks passed, confirming encoding and schema integrity.

**Layer 2 (Semantic Policies):** Semantic checks revealed 10,959 records with negative sales values in state columns—semantically anomalous values representing returns or corrections. A Soda Core check would flag these for steward review. The policy decision requires domain expertise and is escalated to Layer 5.

**Layer 3 (Anomaly Detection):** Statistical monitoring identified a 15.6% decline in glyphosate sales between 2022 (614,329 tonnes) and 2023 (517,983 tonnes). Steward review classified this as natural behavior attributable to documented regulatory changes. A structural change was also identified: two semesters per year for 2007–2021 but only one from 2022 onwards—a schema evolution not documented in the metadata, representing a provenance gap flagged by Layer 4.

**Layers 4–5 (Provenance and Governance):** The IBAMA portal records a last-update timestamp but does not publish a transformation lineage or quality check history. The framework prescribes three governance decisions requiring steward action: (i) classification of 10,959 negative-value records; (ii) documentation of the 2022 semester-structure change; and (iii) establishment of a provenance publication policy for future updates.

## V. Comparative Analysis

We position our framework within existing theoretical landscapes by comparing it to established models across three dimensions: quality focus, governance integration, and operational specificity.

### A. Comparison with Quality Frameworks

ISO/IEC 25012 [3] and Wang and Strong's framework [9] provide comprehensive quality taxonomies but lack operational guidance on implementing continuous validation. Our framework extends these models by mapping quality dimensions to specific validation mechanisms organized hierarchically. Where ISO 25012 defines accuracy as a quality characteristic, our framework specifies structural contracts, semantic policies, and anomaly detection as complementary mechanisms for ensuring accuracy at different levels of abstraction.

Vetrò et al.'s open data quality metrics [4] emphasize measurement but do not address prevention or continuous monitoring. Our framework integrates measurement with proactive validation, shifting from reactive quality assessment to preventive quality engineering.

### B. Comparison with Data Trust Models

Data trust frameworks [5], [6], [8] emphasize governance structures, stakeholder participation, and transparency but provide limited technical specificity regarding quality assurance mechanisms. Our framework operationalizes trust principles by connecting governance decisions to concrete quality validation layers. Where Milne and Brayne emphasize independent stewardship and transparent decision-making [5], our framework specifies how observability and provenance mechanisms enable transparency, and how governance layers translate quality signals into publication decisions.

Recent governance frameworks [7], [29], [30] propose organizational structures and policy guidelines but lack integration with technical quality mechanisms. Our framework bridges this gap by explicitly linking governance decisions to quality assurance outputs.

TABLE I. Comparison of the Proposed Framework Against Related Models.

| Dimension | Proposed | ISO 25012 | FAIR | W3C DQV | Deequ |
|---|---|---|---|---|---|
| Quality dims. | Yes | Yes | Partial | Partial | Yes |
| Exec. mechs. | Yes | No | No | No | Yes |
| Governance | Yes | No | No | No | No |
| Provenance | Yes | No | Partial | Yes | No |
| Anomaly det. | Yes | No | No | No | Yes |
| Open data | Yes | No | Yes | Yes | No |

### C. Comparison with FAIR Principles

FAIR principles [16] provide high-level guidelines for data stewardship but require operationalization through specific mechanisms. Our framework operationalizes FAIR principles: structural contracts ensure interoperability through standardized schemas, provenance tracking enhances findability and reusability, and observability mechanisms support accessibility through transparent quality reporting. Our framework extends FAIR by adding continuous validation and anomaly detection, addressing temporal quality dimensions not explicitly covered by FAIR principles. This extension is critical for open data systems where quality degrades over time through schema drift and distributional shifts.

### D. Gaps in Existing Frameworks

Comparative analysis reveals three critical gaps: (1) fragmentation between quality measurement and governance structures; (2) lack of operational guidance on implementing continuous validation; and (3) insufficient attention to temporal quality dimensions and quality degradation over time. Our framework addresses these gaps by integrating quality assurance with governance, providing hierarchical organization of validation mechanisms, and emphasizing continuous monitoring through anomaly detection and observability. Table I summarises the comparison; our framework is the only model that simultaneously addresses all six dimensions.

## VI. Research Propositions

We articulate six testable propositions linking framework adoption to trust outcomes, quality improvements, and organizational practices. These propositions guide future empirical investigation.

**P1 (Trust and Transparency):** *Open data systems implementing observability and provenance mechanisms (Layer 4) will exhibit higher stakeholder trust compared to systems without such mechanisms, mediated by perceived transparency.*
This proposition draws from data trust literature emphasizing transparency as a trust antecedent [5], [8]. Empirical testing requires measuring stakeholder trust before and after implementing observability mechanisms, controlling for data quality levels.

**P2 (Quality and Validation Layers):** *Open data systems implementing multiple validation layers will demonstrate fewer quality defects in production compared to systems implementing single-layer validation, with diminishing returns beyond three layers.*
This proposition reflects the hierarchical nature of quality assurance, where each layer addresses distinct defect types. Empirical testing requires longitudinal tracking of defect rates across systems with varying numbers of validation layers.

**P3 (Governance and Quality Signals):** *Open data systems integrating quality signals into publication decisions (Layer 5) will exhibit more consistent quality levels over time compared to systems with decoupled quality assessment and publication processes.*
This proposition addresses the gap between quality measurement and governance action. Testing requires comparing quality variance over time between systems with integrated versus decoupled governance.

**P4 (Anomaly Detection and Timeliness):** *Open data systems implementing continuous anomaly detection (Layer 3) will identify quality degradation earlier than systems relying solely on rule-based validation, reducing mean time to detection by at least 50%.*
This proposition emphasizes the value of behavioral monitoring beyond static rules. Testing requires measuring the time elapsed between quality degradation onset and detection across different validation approaches.

**P5 (Semantic Policies and Domain Expertise):** *The effectiveness of semantic policy layers (Layer 2) in preventing quality defects is positively moderated by the level of domain expertise involved in policy definition.*
This proposition recognizes that semantic validation quality depends on domain knowledge. Testing requires comparing defect rates across systems with varying levels of domain expert involvement in policy definition.

**P6 (Framework Adoption and Organizational Maturity):** *Organizations with higher data governance maturity will adopt framework layers in hierarchical order (Layers 1–5), while organizations with lower maturity will adopt layers opportunistically, resulting in lower overall effectiveness.*
This proposition addresses implementation pathways and organizational readiness. Testing requires longitudinal case studies tracking adoption patterns and effectiveness across organizations with varying maturity levels.

## VII. Discussion

This section examines the theoretical, practical, and boundary implications of the proposed framework, situating its contributions within the broader literature and identifying conditions that shape its applicability.

### A. Theoretical Implications

Our framework contributes to data quality theory by integrating quality dimensions with executable validation mechanisms, thereby addressing the gap between measurement and engineering. By organizing mechanisms hierarchically, we provide conceptual clarity about how different validation approaches complement one another. The framework extends software engineering validation practices to open data contexts, demonstrating how test-driven development, design by contract, and observability principles apply to data quality assurance.

The framework also contributes a theoretical account of why existing integrations—DAMA-DMBOK, DataOps literature, and data mesh architectures—fall short for open data systems. DAMA-DMBOK [36] addresses process maturity but not the technical architecture of validation layers. DataOps focuses on pipeline velocity rather than publication governance. Data mesh distributes ownership but does not specify inter-domain quality contracts. Our framework addresses these gaps by providing an explicit mapping from quality signals to governance decisions.

The framework also contributes to trust theory by demonstrating that trust is a systemic property emerging from the interaction of technical mechanisms (Layers 1–4) and organizational processes (Layer 5). Trust cannot be achieved by technical means alone; it requires governance structures that translate quality evidence into accountable decisions. This account bridges the gap between socio-technical trust models and technical quality practices.

The framework also contributes to data trust theory by operationalizing trust principles through technical mechanisms. Where existing trust models emphasize governance structures and stakeholder participation, our framework specifies how quality assurance mechanisms enable transparency, accountability, and informed trust decisions. This integration bridges the gap between socio-technical trust models and technical quality practices.

The framework can be mapped to the three trust dimensions identified by Mayer et al. [43]: *competence* trust (the belief that the trustee has the ability to perform as expected), *integrity* trust (adherence to acceptable principles), and *benevolence* trust (acting in the trustor's interest). Layers 1–3 address competence trust by providing verifiable evidence of correct quality-check performance. Layer 4 addresses integrity trust by making transformation history and quality decisions transparent and auditable. Layer 5 addresses benevolence trust by establishing accountable roles and publication policies that demonstrate the data owner acts in consumers' interests—confirming the framework addresses all three trust dimensions, not merely technical quality assurance.

### B. Practical Implications

For practitioners, the framework provides a roadmap for implementing quality assurance in open data systems. The hierarchical organization suggests implementation priorities: establish structural contracts first, then add semantic policies, followed by anomaly detection and observability. This staged approach enables incremental adoption aligned with organizational maturity and resource constraints.

The framework also guides tool selection and architectural decisions. Each layer maps to specific technology categories: schema validation tools for structural contracts [23], rule engines for semantic policies, statistical monitoring for anomaly detection, and provenance systems for observability. This mapping helps practitioners translate the conceptual framework into concrete implementations.

### C. Limitations and Boundary Conditions

Our framework is conceptual and requires empirical validation through case studies and controlled experiments. The propositions articulated in Section VI provide starting points for such validation but remain untested. The framework emphasizes technical quality mechanisms and may underweight social and organizational factors that influence trust in practice. Trust is fundamentally socio-technical, emerging from interactions among technical systems, organizational practices, and stakeholder relationships. While our framework addresses the technical mechanisms that enable transparency and accountability, it does not fully specify the organizational processes and stakeholder engagement practices required for trust building.

The framework assumes that organizations have sufficient technical capacity to implement all validation layers. For resource-constrained organizations, full implementation may be infeasible. Future work should investigate lightweight implementations and identify minimum viable configurations for different organizational contexts. Trade-offs between validation rigor and computational cost require investigation, particularly for high-volume data streams where comprehensive validation may introduce unacceptable latency [19]. Incremental adoption pathways that deliver value at each stage while building toward comprehensive implementation need further specification.

The framework focuses on structured and semi-structured data [18], with limited applicability to unstructured data (text, images, video). Extending the framework to unstructured data contexts requires additional conceptual development, particularly for semantic validation and anomaly detection layers. Quality dimensions for unstructured data differ from those for structured data, emphasizing aspects such as relevance, coherence, and contextual appropriateness that resist formal specification. Machine learning-based quality assessment for unstructured data introduces additional challenges, including model bias and interpretability.

The framework does not explicitly address adversarial scenarios in which data producers intentionally manipulate quality metrics or validation mechanisms. Security considerations—including data integrity verification, tamper detection, and audit trail protection—require integration with the framework. The relationship between data quality assurance and data security practices merits further investigation.

## VIII. Conclusion and Future Work

This paper proposed a five-layer conceptual framework for trust engineering in open data systems, integrating quality assurance mechanisms with governance principles. The framework organizes validation mechanisms hierarchically—from structural contracts to semantic policies, anomaly detection, observability, and governance—with each layer addressing distinct quality dimensions while collectively building trust through continuous validation and transparent reporting.

Through comparative analysis, we positioned the framework within existing theoretical landscapes, including ISO quality standards, data trust models, and FAIR principles, demonstrating how it extends current models by explicitly linking quality dimensions to executable mechanisms and governance decisions. We identified three critical gaps in existing frameworks: fragmentation between quality measurement and governance, lack of operational guidance on continuous validation, and insufficient attention to temporal quality dimensions.

We articulated six testable propositions linking framework adoption to trust outcomes, quality improvements, and organizational practices, providing a research agenda for empirical validation. The framework contributes a conceptual foundation for engineering trustworthy open data systems that balances transparency, risk management, and stakeholder accountability, bridging the gap between data quality theory, data trust governance, and software engineering validation practices.

Future research should pursue four directions. First, empirical validation through longitudinal case studies tracking framework adoption, implementation challenges, and trust outcomes across diverse organizational contexts. Such studies should test the six propositions in Section VI, examining both successful implementations and failed adoption attempts to identify critical success factors, and compare adaptations across domains (government, healthcare, scientific research).

Second, development of a reference implementation as an open-source pipeline integrating Great Expectations [38], Soda Core [39], dbt [40], and a W3C PROV-DM provenance store, deployed against the IBAMA dataset as a reproducible benchmark. Reference architectures for common technology stacks (cloud platforms, data lakes, data meshes [35]) would provide concrete implementation guidance.

Third, extension of the framework to emerging data contexts, including real-time streaming data, federated data systems, and artificial intelligence training datasets. Streaming contexts require adaptation of anomaly detection mechanisms to handle concept drift and temporal dependencies. Federated systems require distributed validation and consistency protocols across organizational boundaries. AI training datasets require additional validation layers addressing bias, represen-

tativeness, and fairness [33], including training-serving skew and model decay [34].

Fourth, investigation of organizational factors influencing framework adoption and effectiveness. Research should examine how organizational culture, governance maturity, resource constraints, and stakeholder diversity affect implementation success and trust outcomes. The relationship between framework adoption and broader DataOps practices [31] merits exploration, as does integration with data mesh architectures emphasizing domain ownership and federated governance [35].

## REFERENCES


[1] B. W. Wirtz, J. C. Weyerer, and C. Geyer, "Artificial Intelligence and the Public Sector—Applications and Challenges," *International Journal of Public Administration*, vol. 42, no. 7, pp. 596–615, 2019.

[2] J. Attard, F. Orlandi, S. Scerri, and S. Auer, "A systematic review of open government data initiatives," *Government Information Quarterly*, vol. 32, no. 4, pp. 399–418, 2015.

[3] ISO/IEC, *ISO/IEC 25012:2008—Software Engineering—Software Product Quality Requirements and Evaluation (SQuaRE)—Data Quality Model*. Geneva: International Organization for Standardization, 2008.

[4] A. Vetrò, L. Canova, M. Torchiano, C. O. Minotas, R. Iemma, and F. Morando, "Open data quality measurement framework: Definition and application to open government data," *Government Information Quarterly*, vol. 33, no. 2, pp. 325–337, 2016.

[5] R. Milne, J. Morley, and H. S. Howard, "Data trusts and the governance of health data," *Nature Medicine*, vol. 28, pp. 2218–2220, 2022.

[6] D. Radosevic, M. Cetl, and V. Cetl, "Spatial Data Trust Framework," *ISPRS International Journal of Geo-Information*, vol. 12, no. 11, p. 456, 2023.

[7] B. C. Stahl, D. Wright, and M. Wakunuma, "Ethics of AI and big data: Governance frameworks and their implications," *AI & Society*, vol. 40, pp. 1–15, 2025.

[8] A. Artyushina, "The civic data trust: A new model for data stewardship," *Data & Policy*, vol. 2, e7, 2020.

[9] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996.

[10] K. Beck, *Test-Driven Development: By Example*. Boston, MA: Addison-Wesley, 2003.

[11] B. Meyer, "Applying 'Design by Contract'," *Computer*, vol. 25, no. 10, pp. 40–51, 1992.

[12] C. Majors, L. Fong-Jones, and G. Sheffield, *Observability Engineering*. Sebastopol, CA: O'Reilly Media, 2022.

[13] J. Cheney, L. Chiticariu, and W. C. Tan, "Provenance in databases: Why, how, and where," *Foundations and Trends in Databases*, vol. 1, no. 4, pp. 379–474, 2009.

[14] Y. L. Simmhan, B. Plale, and D. J. Gannon, "A survey of data provenance in e-Science," *ACM SIGMOD Record*, vol. 34, no. 3, pp. 31–36, 2005.

[15] P. Groth and L. Moreau, "PROV-Overview: An Overview of the PROV Family of Documents," W3C Working Group Note, Apr. 2013. [Online]. Available: https://www.w3.org/TR/prov-overview/ [retrieved: Apr. 2026].

[16] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, p. 160018, 2016.

[17] N. Nicholson, R. N. Carvalho, and I. Šotl, "A FAIR Perspective on Data Quality Frameworks," *Data*, vol. 10, no. 9, p. 136, Aug. 2025.

[18] C. J. Date, *Database Design and Relational Theory*, 2nd ed. Sebastopol, CA: O'Reilly Media, 2019.

[19] M. Kleppmann, *Designing Data-Intensive Applications*. Sebastopol, CA: O'Reilly Media, 2017.

[20] B. W. Boehm, "Verifying and validating software requirements and design specifications," *IEEE Software*, vol. 1, no. 1, pp. 75–88, Jan. 1984.

[21] IEEE, *IEEE Standard for System, Software, and Hardware Verification and Validation*, IEEE Std 1012-2016. New York, NY: IEEE, 2017.

[22] L. Getoor and A. Machanavajjhala, "Entity resolution: Theory, practice & open challenges," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2018–2019, 2012.

[23] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Grafberger, "Automating large-scale data quality verification," *Proceedings of the VLDB Endowment*, vol. 11, no. 12, pp. 1781–1794, 2018.

[24] Brazilian Institute of Environment and Renewable Natural Resources (IBAMA), "Comercialização de Agrotóxicos por Unidade da Federação," Open Data Portal, 2024. [Online]. Available: https://dadosabertos.ibama.gov.br/dados/AGROTX/produtosestado_csv.zip [retrieved: Apr. 2026].

[25] OpenAPI Initiative, "OpenAPI Specification v3.2.0," 2025. [Online]. Available: https://spec.openapis.org/oas/latest.html [retrieved: Apr. 2026].

[26] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.

[27] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA: Morgan Kaufmann, 2011.

[28] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "MOA: Massive online analysis," *Journal of Machine Learning Research*, vol. 11, pp. 1601–1604, 2010.

[29] A. Sarran, R. Ramnarain-Seetohul, and S. Cadersaib, "Towards a Data Governance Model for Enhanced Data Quality Management," *International Journal of Information Technology, Research and Applications*, vol. 3, no. 4, p. 110, 2024.

[30] O. Adeyemi and C. Okonkwo, "A conceptual framework for data governance in big data and cloud environments," *International Journal of Science and Research Archive*, vol. 12, no. 2, pp. 1177–1195, 2024.

[31] A. Reis and R. Housley, "DataOps: Towards a Definition," in *Proc. 13th Int. Conf. Software Technologies (ICSOFT)*, Porto, Portugal, 2018, pp. 104–111.

[32] Food Standards Agency, "Food Data Trust: Legal, Structuring and Governance," *FSA Research and Evidence*, London, UK, 2021.

[33] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, MA: MIT Press, 2023.

[34] D. Sculley *et al.*, "Hidden technical debt in machine learning systems," in *Proc. 29th Conf. Neural Information Processing Systems (NIPS)*, Montreal, Canada, 2015, pp. 2503–2511.

[35] Z. Dehghani, *Data Mesh: Delivering Data-Driven Value at Scale*. Sebastopol, CA: O'Reilly Media, 2022.

[36] DAMA International, *DAMA-DMBOK: Data Management Body of Knowledge*, 2nd ed. Basking Ridge, NJ: Technics Publications, 2017.

[37] R. Albertoni and A. Isaac, "Data on the Web Best Practices: Data Quality Vocabulary," W3C Working Group Note, Dec. 2016. [Online]. Available: https://www.w3.org/TR/vocab-dqv/ [retrieved: Apr. 2026].

[38] Great Expectations, "Great Expectations: Data Validation Framework," 2023. [Online]. Available: https://greatexpectations.io [retrieved: Apr. 2026].

[39] Soda, "Soda Core: Open-Source Data Quality Framework," 2023. [Online]. Available: https://github.com/sodadata/soda-core [retrieved: Apr. 2026].

[40] dbt Labs, "dbt: Data Build Tool," 2023. [Online]. Available: https://www.getdbt.com [retrieved: Apr. 2026].

[41] J. Wu, Y. Tian, F. Thung, D. Lo, and C. Tantithamthavorn, "Data Quality Matters: A Case Study on Data Label Correctness for Security Bug Report Prediction," *IEEE Transactions on Software Engineering*, vol. 48, no. 7, pp. 2541–2556, Jul. 2021.

[42] X. Gong, J. Liu, Y. Zhao, and H. Wang, "A survey on dataset quality in machine learning," *Information and Software Technology*, vol. 162, p. 107268, Oct. 2023.

[43] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Academy of Management Review*, vol. 20, no. 3, pp. 709–734, 1995.