April 2025

A Data Quality and Data Ethics Framework

for AI Readiness in Kenya

Report by

Tech Innovators Network

(THiNK) Tank

Table of contents

Foreword 1

Why We Need a Data Quality Framework 2

Principles of the DQF 3

How to Use the Data Quality Framework 4

The Six Dimensions of Data Quality 8

Next Steps: Advancing the DQF 11

Data Quality Maturity Model Framework 12

Annex A : Supporting Research and Outputs

Foreword

In this digital age, the power of data is not just in its volume, but in its conformity to facts. From informing public health decisions to enabling inclusive digital services, data is the cornerstone of informed governance, citizen trust, and technological progress. However, without integrity, data becomes a liability rather than an asset.

In response to the critical need for a unified, practical guide for data governance, we at the **Tech Innovators Network (THiNK) Tank**, initiated the development of the **Data Quality Framework (DQF)**—a flexible, actionable, and context-aware framework designed to help institutions assess, manage, and improve data quality across sectors. Grounded in both global standards and local realities, the DQF was shaped through institutional partnerships, pilot evaluations, and feedback.

We now extend this framework as a strategic foundation for institutional data quality transformation. The findings from pilot audits, using datasets from Strathmore University, and field assessments confirm the urgent need for stronger documentation, standardized validation protocols, and governance structures.

This report presents the full scope of the framework, complemented by lessons learned during implementation, and clear recommendations for continuous improvement. The work was graciously sponsored by UK International Development, through the UK-Kenya AI Challenge, a program administered through the Africa Center of Technology Studies (ACTS).

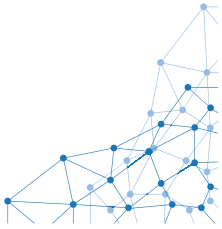
Why Do We Need a Data Quality Framework ?

Despite Kenya's growing reliance on data to power digital services, planning, and public accountability, significant quality issues persist. Some of the challenges highlighted that established the need of a clear data quality framework for better AI data included:

- **Unclear Data Provenance** - It is often difficult to establish who the real owner and source of the data is. For instance, after evaluation of the farmer datasets we discovered that the enumerators often did not clearly define whether respondents are farmers, farm workers, or other stakeholders—compromising contextual accuracy.
- **Inconsistent Validation** - Researchers use personal methods to validate data without centralized documentation or standardization.
- **Lack of Formal Agreements with Third-Party Data Providers,** often relying on downloaded text files without APIs or service-level assurances.
- **Regulatory reliance** - Projects often defer to university compliance instead of maintaining dedicated project-level regulatory audits.
- **Missing documentation for data cleaning, transformation, or verification process**

A **DQF addresses these systemic gaps** by creating a national level standard and toolset for:

Establishing data verification procedures

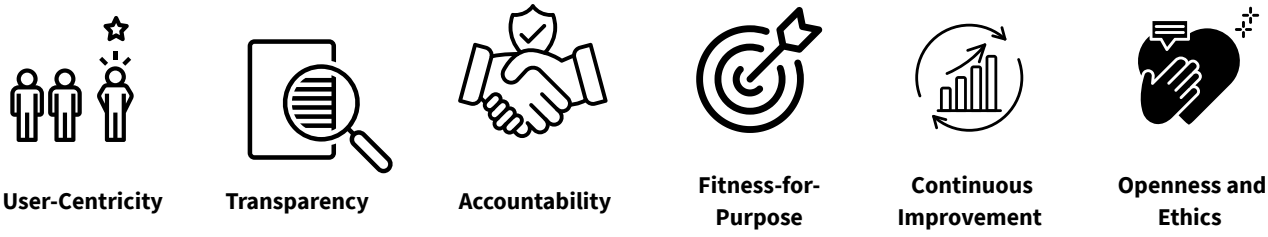


Creating compliance-ready documentation

Promoting inter-agency interoperability

Principles of the DQF

The Data Quality Framework shall seek to establish a series of tools, processes and procedures that can be used to achieve assurance of data quality. The principles provide a high level target of quality or ethics, that the framework wishes to achieve upon application. The prinicples are generally more agile, and can be established in consensus with users and/or pracitioners. Upon review of the Report of the IC&DE Sectoral Working Group (2024), the Kenya National AI Strategy (2025), the Kenya National AI Principles (2025) and the (Draft) Kenya AI Code of Practice Standard (2025), we established the following data principles:



- **User-Centricity** - Data must meet the needs of its users.
- **Transparency** - Metadata, lineage, and transformations must be traceable.
- **Accountability** - Roles and responsibilities must be clearly defined and enforced.
- **Fitness-for-Purpose** - Data should be suitable for the context in which it is used.
- **Continuous Improvement** - Regular updates and audits are mandatory
- **Openness & Ethics** - Data must respect privacy, equity, and openness.

Practical data initiatives continue to highlight the importance of structured quality frameworks.

For example, the Mozilla Common Voice project, to which THiNK contributed as a local partner, demonstrates the importance of community-driven, multilingual dataset development. This effort focused on collecting high-quality voice data in Swahili and indigenous languages, reflects key DQF principles—especially openness, relevance, and ethical data collection. Similarly, Dr. Betsy Muriuki's agricultural data research uncovered insight into how data fragmentation and undocumented validation methods limit the use of farmer-level data in planning and intervention. The need for a DQF would guarantee easier institutionalizing of ethical data practices and privacy protections.

How to Use the Data Quality Framework

The Data Quality Framework (DQF) is designed for both flexibility and structure. It can be adopted across different institutional scales—from small data units to large ministries—while ensuring that all users benefit from a unified approach to data quality improvement.

To facilitate successful adoption, the DQF outlines a **Six-Phase Implementation Pathway**. Each phase represents a logical step in institutionalizing high-quality data practices, starting from internal assessment and culminating in system-wide integration and knowledge sharing.

- **PHASE 1: READINESS & BASELINE ASSESSMENT**
- **PHASE 2: DEFINE QUALITY OBJECTIVES**
- **PHASE 3: DEVELOP DATA QUALITY MODEL**
- **ASSIGN STEWARDSHIP AND GOVERNANCE ROLES**
- **PHASE 4: CAPACITY BUILDING AND TOOL DEPLOYMENT**
- **PHASE 5: CONDUCT PILOTS AND REFINE TOOLS**
- **PHASE 6: MONITORING & DOCUMENTATION & KNOWLEDGE SHARING**

Phase 1: Readiness and Baseline Assessment

This initial phase establishes the foundational understanding of an institution's current data landscape.

Key Actions

- **Map core datasets**

  Identify datasets that are mission-critical, frequently used, or regulatory in nature.

- **Use the DQF Questionnaire and Audit Template**

  Assess current practices across data sourcing, verification, preparation, consent, privacy, and stewardship.

- **Diagnose challenges**

  Highlight inconsistencies, undocumented processes, lack of validation tools, or incomplete consent and privacy frameworks.

- **Categorize datasets by risk and sensitivity**

  Flag those requiring urgent quality interventions (e.g., datasets related to health, education, finance, or vulnerable populations).

Outcome

A comprehensive baseline report that informs strategic prioritization in Phase 2.

Phase 2: Define Quality Objectives

In this phase, institutions tailor the DQF principles and quality dimensions to their operational context and strategic goals.

Key Actions

- **Select quality indicators** relevant to your datasets (e.g., accuracy rate for survey data, timeliness for emergency response data).
- **Set realistic benchmarks** for improvement (e.g., reduce missing fields to <5% in the next quarter).
- **Align with mandates and policies** - Ensure data goals support legal obligations (Data Protection Act, sectoral regulations) and institutional mandates (e.g., service delivery targets, audit requirements).
- **Differentiate by dataset type** - Define objectives differently for operational datasets (requiring precision) and strategic datasets (requiring relevance).

Outcome

Customized quality improvement plans with clear, measurable goals.

Phase 3: Assign Stewardship and Governance Roles

Effective data quality management requires clear accountability and distributed leadership.

Key Actions

- Designate data stewards for each major dataset or department. These individuals are responsible for ensuring quality, documentation, and compliance.
- Define roles and responsibilities clearly in data governance policies.
- Establish or strengthen a Data Governance Committee, chaired by a senior leader, to coordinate implementation across departments.
- Empower stewards with authority and resources to lead quality audits, training, and remediation efforts.

Outcome

A governance structure that ensures continuity, oversight, and accountability in quality assurance.

Phase 4: Capacity Building and Tool Deployment

This phase focuses on equipping staff and systems with the technical and human capabilities required to maintain data quality.

Key Actions

- Train institutional staff and stewards using the DQF training manual, online learning modules, and facilitated workshops.
- Deploy validation tools and scripts in data pipelines (e.g., automated checks for missing values, duplications, or outliers using tools like R or Python).
- Customize templates for metadata documentation, consent forms, data verification logs, and quality scorecards.
- Establish feedback loops for teams to report issues, suggest improvements, and monitor progress over time.

Outcome

A technically equipped workforce and an operational system capable of real-time and retrospective data validation.

Phase 5: Conduct Pilots & Refine

Testing the framework in real operational environments helps reveal practical insights and refine approaches before full-scale rollout.

Key Actions

- Select high-priority sectors for piloting—such as agriculture, health, education, or social protection.
- Implement the DQF across a full data lifecycle, from collection and preprocessing to use and archiving.
- Document pilot findings, including data quality improvements, challenges faced, and feedback from users and data consumers.
- Refine tools and procedures based on real-world experiences (e.g., simplifying checklists, adjusting frequency of audits).

Outcome

Field-tested improvements and validated adaptations that make the framework more robust and scalable.

Phase 6: Documentation and Knowledge Sharing

Capturing and sharing institutional learning is vital for building a culture of continuous improvement and sectoral coherence.

Key Actions

- Publish institutional data quality reports or case studies for internal governance and public accountability.
- Share success stories and challenges at cross-sectoral forums, peer exchanges, or with the NDQAC (National Data Quality Assurance Committee).
- Institutionalize lessons learned by updating SOPs, integrating new practices into onboarding and training, and revising data management policies.

Outcome

A knowledge-driven culture where quality improvement is iterative, evidence-based, and widely disseminated.

The DQF outlines six key dimensions of data quality. Each dimension captures a vital aspect of data performance and reliability. The combined application of these dimensions allows institutions to diagnose, monitor, and improve the quality of data in a systematic way.

1. Accuracy

Definition: The extent to which data correctly reflects the real-world values or events it is intended to represent.

Why does it matter? Inaccurate data can lead to misinformed decisions, policy failures, and mistrust from the public. Indicators:

- Error rates (e.g., incorrect ID numbers, misspelled names)
- Validation against source data
- Frequency of manual correction

2. Completeness

Definition: The extent to which all required data is present and recorded without gaps.

Why does it matter ? Incomplete data leads to poor analytical outcomes and underrepresentation. In public health, for instance, missing vaccination data could affect outbreak predictions and response.

Indicators

- Percentage of blank or null fields
- Missing records in time-series datasets

3. Consistency

Definition: The extent to which data is presented in a uniform and logical format across datasets and systems.

Why it matters - Inconsistent data leads to duplication, inefficiencies, and conflict in policy implementation.

Indicators

- Use of standard formats (e.g., date/time)
- Duplication rates
- Alignment across interoperable systems

4. Timeliness

Definition: The extent to which data is available and up-to-date when needed.

Why it matters

Late data leads to missed opportunities. Real-time or near-real-time data is critical in fast-moving sectors like agriculture (e.g., rainfall data), health (e.g., disease outbreaks), or emergency response.

Indicators

- Time lag between event and data entry/publication
- Data refresh rates
- Availability of real-time feeds or APIs

5. Accessibility

Definition: The ease with which data can be retrieved and used by authorized stakeholders.

Why it matters - Even high-quality data loses value if it's locked away or requires excessive effort to access. Accessibility is essential for transparency.

Indicators

- Availability through APIs, dashboards, or open portals
- Licensing and reuse policies
- Documentation and metadata availability

6. Relevance

Definition: The extent to which data meets the needs of users and supports its intended purpose.

Why it matters -Collecting high volumes of data is not enough—it must serve a clear purpose. Irrelevant data wastes resources and leads to analytic overload.

Indicators

- Stakeholder satisfaction surveys
- Alignment with decision-making requirements
- Demand vs. usage metrics.

Next Steps: Advancing the DQF

1. Standardize Third-Party Data Agreements

- Develop API-based data sourcing with embedded Service Level Agreements(SLAs).

2. Intergrate a TFGBV Lexicon .

- Integrate machine-readable terms into DQF-compliant datasets to support real-time detection.

3. Train Institutional Stewards

- Establish certification pathways in collaboration with Strathmore and Maseno.

4. Embed DQF in Digital Platforms

- Possibly align the framework with e-citizen and GovStack architectures

Next Steps: Advancing the DQF

1. Standardize Third-Party Data Agreements

- Develop API-based data sourcing with embedded Service Level Agreements(SLAs).

2. Intergrate a TFGBV Lexicon .

- Integrate machine-readable terms into DQF-compliant datasets to support real-time detection.

3. Train Institutional Stewards

- Establish certification pathways in collaboration with Strathmore and Maseno.

4. Embed DQF in Digital Platforms

- Possibly align the framework with e-citizen and GovStack architectures

Data Quality Maturity Model Framework

Objective:

Ensure that AI initiatives have the needed data for effective and successful implementation. It is designed to guide organizations in developing the robust data capabilities essential for successful, ethical, and scalable Artificial Intelligence (AI) initiatives. The provided foundational framework, structured around the four pillars of Architecture, Security & Privacy, AI Data Governance, and Accessibility, has been systematically reviewed and enriched by integrating a comprehensive suite of internationally recognized standards and best practices. This transformation elevates the model from a high-level conceptual guide to a strategic, actionable, and auditable roadmap for achieving enterprise-wide data maturity.

This journey involves establishing foundational governance, implementing managed processes, and ultimately leveraging data as a strategic asset for competitive advantage. Adopting this structured, standards-based approach is a business imperative, enabling organizations to de-risk complex AI investments, accelerate innovation, and build lasting trust with customers, regulators, and the public. Performance Outcomes: Data is available, accessible, and secure for the development and operation of AI capabilities.

Key Performance Indicators: Indicators are measures for: AI data security, volume, governance, accessibility, variety, velocity, and veracity.

The Data Pillar has four dimensions

1. **Architecture**
2. **Security and Privacy**
3. **AI Data Governance**
4. **Accessibility**

I. Data architecture Maturity model

Enterprise Data Architecture is the practice of defining the business's data strategy and designing the blueprints for managing data assets. It encompasses models, policies, rules, and standards that govern which data is collected, how it is stored, arranged, integrated, and put to use in data systems and in organizations, with a specific focus on supporting scalable, reliable, and ethical AI initiatives. This moves beyond simple data storage to a strategic function that enables business objectives through well-managed data.

Level 1

No common AI data architecture/framework is in place.

At this initial level, no formal AI data architecture exists. Data is created, stored, and managed within disconnected application silos, leading to widespread redundancy and inconsistency. There is no common business lexicon or enterprise-wide data ontology, meaning the same concept (e.g., "customer") may be defined differently across departments. Data required for AI projects is sourced through ad-hoc, often manual and time-consuming, extraction processes.

Level 2

An initial AI data architecture/framework is being developed.

The organization recognizes the limitations of the ad-hoc approach, typically driven by the needs of a single, high-priority AI initiative. An initial AI data architecture is being developed, though its scope may be limited to that specific project. Efforts begin to identify and catalog key data assets and to create a preliminary business glossary to standardize critical terms.

Standards: KS 3007:2025, KS ISO/IEC/IEEE 42010

Level 3

An approved enterprise-wide data architecture/framework is consistent with the AI implementation plan and the needs of AI initiatives.

A comprehensive, enterprise-wide data architecture is formally defined, documented, and approved by executive stakeholders. This architecture includes a standardized enterprise data model, a managed business lexicon and ontology, and clear principles for data storage, integration, and lifecycle management. Crucially, the architecture is designed to explicitly support the needs of AI, incorporating requirements for data provenance, version control for datasets, and structures that facilitate model training and validation.

Standards: KS 3007:2025, ISO 8000-100 Series, TOGAF

Level 4

An enterprise-wide data architecture/framework is implemented, consistently utilized, and monitored.

The enterprise data architecture is no longer just a blueprint; it is fully implemented and operational across the organization. Compliance with architectural standards is actively managed and enforced through governance processes. A set of key performance indicators (KPIs) is used to continuously monitor the architecture's effectiveness. For AI, these metrics might include the reduction in data preparation time for data scientists, the speed of provisioning new data environments, or the reusability of data assets across multiple AI models.

Standards: KS 3007:2025, ISO 8000-63, TOGAF: Phase G, ISO 19650.

Level 5

An optimized common data architecture/framework is utilized across the enterprise and is updated to meet evolving needs of AI initiatives

The data architecture transcends a static, managed state and becomes a dynamic, agile system. It is continuously improved and optimized based on performance metrics, evolving AI requirements, and emerging technological paradigms. The architecture is designed for adaptability, potentially incorporating advanced concepts like data mesh or data fabric.

This allows for the decentralization of data ownership to domain-specific teams while maintaining strong central governance through a shared, self-service data platform.

Standards: KS 3007:2025, ISO/IEC AWI 20151, TOGAF: Architecture Development Method

2. Security and Privacy Model

Protection of privacy rights and data security rights for AI is embedded and upheld by individuals designing, using, and overseeing AI systems to control the safety, specificity, and exchange of personal digital information.

Security and Privacy in the context of an AI data maturity model refers to the systematic implementation, operation, and continual improvement of a comprehensive set of controls designed to protect the confidentiality, integrity, and availability of all data assets used in AI initiatives.
It explicitly includes the governance required to ensure the lawful, fair, and ethical processing of Personally Identifiable Information (PII) in alignment with global privacy regulations and the Data Protection Act 2019.

Level 1

No AI-specific privacy and data security rights, approaches, and standards are in place for individual security, control, safety, and specificity in exchange for digital information.

There are no formal, documented, or AI-specific security and privacy policies. Security measures are implemented reactively, typically in response to a specific threat or incident. Basic perimeter controls like firewalls may be in place, but there is no overarching Information Security Management System (ISMS) to ensure controls are comprehensive, consistently applied, or effective. Privacy considerations are an afterthought.

Standards Mapping: No formal standards are applied.

Level 2

AI-specific privacy and data security rights, approaches, and standards for individual control of safety, specificity, and exchange of digital information.

The organization recognizes the need for a more structured approach. Initial AI-specific security and privacy policies are being drafted, often driven by legal and compliance requirements. There is an awareness of major regulations like GDPR and DPA, and a preliminary effort is made to identify key security risks associated with AI systems. Some security controls, perhaps drawn from a recognized standard, are implemented, but this is often done in an uncoordinated, project-specific manner without a formal risk assessment to justify their selection

Standards: ISO/IEC 27001, Data protection Regulation

Level 3

AI-specific privacy and data security rights, approaches, and standards defined and approved for individual control of safety, specificity, and exchange of digital information.

A formal, enterprise-wide Information Security Management System (ISMS) is designed and approved, with its scope, policies, and objectives explicitly aligned with the ISO/IEC 27001 standard.
A documented risk assessment methodology is established and used to systematically identify and evaluate information security risks. Based on this assessment, a comprehensive set of security policies and controls is defined and approved.

For privacy, a Privacy Information Management System (PIMS), aligned with ISO/IEC 27701, is designed as a formal extension to the ISMS.

Standards: KS 3007:2025, ISO/IEC 27001, ISO/IEC 27701

Level 4

Enterprise-wide, AI-specific privacy and data security rights, approaches, and standards for individual control of safety, specificity, and exchange of digital information are managed and monitored.

The defined ISMS and PIMS are fully implemented and operational across the enterprise. Security and privacy are no longer just documented policies; they are managed business processes. The effectiveness of controls is continuously monitored, and a program of regular internal audits is in place to verify compliance. Security and privacy metrics are collected, analyzed, and formally reviewed by management.

All relevant staff receive regular security and privacy awareness training, and compliance with policies is actively enforced.

Standards: KS 3007:2025, ISO/IEC 27001, ISO/IEC 27701, NIST Cybersecurity Framework

Level 5

Enterprise-wide, AI-specific privacy and data security rights, approaches, and standards for individual control of safety, specificity, and exchange of digital data are improved and optimized based upon data trends.

The defined ISMS and PIMS are fully implemented and operational across the enterprise. Security and privacy are no longer just documented policies; they are managed business processes. The effectiveness of controls is continuously monitored, and a program of regular internal audits is in place to verify compliance. Security and privacy metrics are collected, analyzed, and formally reviewed by management. All relevant staff receive regular security and privacy awareness training, and compliance with policies is actively enforced.

Standards: KS 3007:2025, ISO/IEC 27001, ISO/IEC 27701, NIST Cybersecurity Framework

3) AI Data Governance

Process of managing AI data performance and compliance to guard against data bias and ensure availability, usability, and integrity of data in AI systems.

AI Data Governance is the overarching framework of authority, control, and decision-making for managing data assets within AI systems. It encompasses the people, processes, and technologies required to ensure that AI-enabling data is managed as a strategic enterprise asset. This includes establishing and enforcing policies for data performance, quality, integrity, security, and compliance.

Critically, its mandate extends beyond traditional data governance to specifically identify, measure, monitor, and mitigate the unique risks posed by AI, such as algorithmic bias, lack of transparency, poor data provenance, and unintended societal impacts. The progression of maturity in AI Data Governance can be effectively mapped to the iterative, four-function core of the NIST AI Risk Management Framework: Govern, Map, Measure, and Manage. An organization's ability to execute these functions determines its maturity level.

Level 1

No AI governance structure, audits, processes, and standards are in place.

No formal AI data governance structure exists. Data ownership is undefined, accountability is absent, and there are no standardized processes for managing data quality or assessing AI models for bias. Decisions about data handling, ethics, and risk are made inconsistently by individual project teams, if at all.

Standards : No governance standards are applied.

Level 2

AI governance and processes to guide and oversee how AI is developed and used are initiated. The organization recognizes the need for AI-specific governance, often prompted by problematic AI deployment or a new regulatory requirement. An initial governance charter is drafted, and informal discussions begin around assigning data-related roles. A preliminary inventory of AI systems and the data they consume is initiated.

Standards: : KS 3007:2025, IEEE P2863, NIST AI RMF, ISO/IEC TS 38505-3:2021

Level 3

AI governance and processes are established, accountable executives are identified, and audits are in place and utilized.

A formal, enterprise-wide AI data governance framework is defined, documented, and approved by executive leadership. A dedicated governance body, such as an AI Governance Council or an AI Ethics Committee, is established with a clear charter and decision-making authority. Key roles and responsibilities (e.g., Data Owners, Data Stewards, AI Risk Managers) are formally defined and assigned. A comprehensive set of policies covering data quality, AI ethics, bias assessment, and model transparency is written and approved.

Standards: KS 3007:2025, NIST AI RMF, ISO 8000-61, IEEE 7005-2021, IEEE P2863, ISO/IEC AWI 25590

Level 4

AI governance is routinely carried out, and participation is representative of the organizational entity. Metrics are consistently collected and inform adherence to defined standards.

The AI data governance framework is fully operational and integrated into the organization's daily activities. Policies are consistently enforced, and the AI Governance Council meets regularly to review new and existing AI initiatives. A robust set of metrics for data quality, model fairness, transparency, and overall governance effectiveness is continuously collected and monitored. Formal audits of high-risk AI systems are conducted periodically to ensure compliance with internal policies and external regulations.

Standards: KS 3007:2025, NIST AI RMF, ISO 8000

Level 5

utilization by re-evaluation of existing standards, processes, policies, and procedures.

The AI data governance framework operates as a dynamic learning system. Insights from continuous monitoring, audits, and incident response are systematically fed back to refine policies, processes, and metrics. The organization actively engages with the broader industry and academic communities on AI ethics and standards, demonstrating leadership. Automated tools are increasingly used to support governance functions, such as continuous model monitoring for drift and bias, and managing the risk management lifecycle

Standards: KS 3007:2025, NIST AI RMF, IEEE Standards

4.Accessibility

A systematic approach and structure to address the challenges, legal agreements, and requirements needed for managing trusted and secure data sharing both internally and externally.
Data Accessibility, in its mature form, transcends the narrow scope of legal data sharing agreements. It is the comprehensive establishment of principles, policies, agreements, and technical infrastructure required to ensure that data assets are Findable, Accessible, Interoperable, and Reusable (FAIR) for both human experts and machine-driven AI systems.

This dimension encompasses not only the procedural and legal frameworks for trusted data sharing but also the critical technical mechanisms for data discovery (e.g., data catalogs), semantic interoperability (e.g., shared ontologies and vocabularies), and machine-actionable access (e.g., standardized APIs). The ultimate goal is to enable a trusted, secure, and efficient flow of data to fuel and scale AI initiatives across the enterprise.

Level 1

No formal data sharing framework or agreements exist.
Data is fundamentally inaccessible, locked within operational application silos. Finding relevant data requires tribal knowledge of systems and people. When data is shared, it is done via informal, point-to-point methods such as email attachments or manual file transfers, with no record, security, or consistency.

Standards : No standards are applied.

Level 2

A data-sharing framework is in process to provide common data-sharing agreements and facilitate data cataloging decoupling (freeing data from applications) and eliminating silos. The organization recognizes that data silos are a major impediment to progress. A basic data sharing framework is being developed, and initial efforts are made to create a rudimentary catalog of key datasets. The first data sharing agreements are drafted, typically for a specific, high-priority project, to formalize an exchange between two departments/organizations. The primary focus is on breaking down the most critical data silos to enable a specific use case.

Standards: FAIR Principles, Data Sharing Agreement Best Practices

Level 3

An approved data-sharing framework and agreements are in place that provide the data availability, accessibility, and quality needed to support AI.
An enterprise-wide data sharing framework, including a standard, legally-vetted template for data sharing agreements, is formally defined and approved.

A central data catalog is established and populated with rich metadata for key enterprise data assets.
Clear policies and procedures are documented, defining how data can be requested, approved, and accessed. The framework is explicitly designed with the goal of making data FAIR.

Standards: FAIR Principles, Data Sharing Agreement Best Practices, W3C DCAT (Data Catalog Vocabulary)

Level 4

A data-sharing framework and agreements are routinely used and verified by metrics to support consistency in data usage.
The data sharing framework is fully operational and routinely used across the enterprise. The automated data catalog is the primary, trusted mechanism for discovering and requesting access to data. Data access is increasingly managed through standardized, secure APIs rather than manual file transfers. A set of metrics is used to track data usage, monitor the efficiency of sharing request fulfillment, and verify compliance with the terms of data sharing agreements.

Standards: FAIR Principles, W3C standards , ISO 19650

Level 5

An enterprise-wide data-sharing service framework and agreements are improved from collected data analytics and best practices.

The data sharing ecosystem is a core strategic capability of the organization and is subject to continuous improvement based on usage analytics and stakeholder feedback. The organization moves beyond internal data sharing and begins to participate in external, federated data ecosystems or "data spaces," securely sharing data with trusted partners to enrich its own AI models and create new value. Data accessibility is viewed not as a support function but as a primary enabler of business innovation.

Standards Mapping: International Data Spaces (IDS), W3C Linked Data Platform (LDP, FAIR Principles

Annex A.2: Case Study – Mozilla Common Voice (Swahili & Indigenous Languages

As a supporting partner, Tech Innovators Network (THiNK) contributed to the Mozilla Common Voice project in Kenya, focusing on the collection and validation of voice samples in Swahili.The goal of the project was to build open-source, publicly accessible datasets that can support more inclusive and linguistically diverse AI systems—especially in voice interfaces and speech recognition.

THiNK supported:

- Community mobilization and recording sessions in underrepresented regions
- Data quality checks and validation feedback loops
- Promoting the ethical use of voice data with informed consent and documentation

This initiative aligns closely with DQF principles such as User-Centricity, Transparency, Openness, and Relevance—proving that ethical, collaborative data projects can meaningfully influence the design of AI systems and public digital tools.

Project Partners & Contributors

This project was executed with input and collaboration from distinguished experts and institutions:

Brian Omwenga

Brian is an AI and Technology Policy Expert. The founder of Tech Innovators Network (THiNK) is focused on the development of a strong and inclusive ICT ecosystem in Africa. After graduating from Strathmore and Massachusetts Institute of Technology (MIT) he has worked fruitfully in the private and the public sector. As a software engineer and project manager at Nokia he successfully started his journey of creating notable inventions and filing patents.
In the public sector he successfully designed a Government Wide Enterprise Architecture covering the business, data, applications and technology domains. As a result of these experiences, he champions the open innovation philosophy through collaboration, inclusiveness and quality output. He is extremely passionate about the role and power of COMMUNITY in the context of the ICT ecosystem.

Dr. Betsy Muriuki – Strathmore University

Dr. Betsy Muriithi is a Research Fellow at @iLabAfrica, where she leverages her expertise in data analytics to address pressing societal challenges. Her research focuses on using artificial intelligence (AI) technologies to drive sustainable development, particularly in agriculture and public health. Her current interests include exploring how these technologies can be used to create practical solutions for communities and businesses. She has developed tools that integrate AI and IoT to support smallholder farmers in enhancing climate resilience and productivity, as well as decision-support systems that improve public health performance monitoring. She also studies human-computer interaction focusing on:
(1) the mechanisms that enable effective data use for decision-making—whether its helping farmers optimize agricultural practices or assisting healthcare managers in boosting facility efficiency, and
(2) how AI and data systems can be designed to be more inclusive, fostering equitable access to technology in communities and businesses.

Monica Okoth – Kenya Bureau of Standards (KEBS)

Monica Okoth Is the Assistant Manager, ICT and Electrotechnical department,Standards development division.She serves as Technical Committee(TC) Manager for several national committees at the Kenya Bureau of Standards (KEBS) on topics related to AI, IT and related technology.
She is also the National Committee Manager for the Joint International Technical Committee for information technology (ISO IEC JTC1). This involves the development of National and international standards on software engineering, Artificial Intelligence,Data Centers, E-Learning, IT service management and IT governance.Monica has participated in developing several national policies including national AI strategy, e-waste regulations, energy efficiency, national communications andaddressing plan and assistive technology and accessibility among others.She is a certified Information Security , Implementor and Auditor and hold a bachelor's degree in mathematics and geography from the University of Nairobi and is working on a postgraduate certificate in applied statistics.

Nick Mumero

Nick is the AI Architect and lead AI Engineer at THiNK, where he oversees AI/ML teams and project delivery. He specializes in multilingual NLP and LLM applications, having led chatbot projects across public and private sectors. At THiNK, he developed the top-performing Swahili ASR model using Wave2vec 2.0 and built multilingual chatbots serving over 15,000 users monthly. He also created THiNKiT, a low-resource smart speaker for English, Swahili, and Kikuyu. Nick has contributed to open-source Swahili models and published research on sentiment analysis, combining deep technical expertise with strong product leadership.

Angela Kanyi

Angela is an experienced software engineer and project manager. At THiNK she currently runs DevOps, where she manages collaborations between the developer team and the users. Angela was integral in the design of the Conformity Assessment Process (CAP) at THiNK. The CAP process ensures that all THiNK projects conform to high quality standards, during the software development process as well as the software-in-use. This has equipped her as the primary liaison between users and developers, ensuring smooth communication and delivering user-centric solutions. Angela has run successful projects having over 5,000+ users while at THiNK. Angela has technical skills in JavaScript, React, DevOps Tools, CI/CD, Cloud Infrastructure, and REST APIs. She possesses excellent experience and knowledge or tools that manage developer operations such as sprints, through to user management tools such as ticketing solutions. Her communication and soft skills are an added asset that have enabled her to be a good problem-solver, communicator, and collaborator.

Conclusion

The Data Quality Framework (DQF) represents a crucial step toward improving data governance across sectors.

The framework addresses systemic data quality gaps through practical, flexible, and context-aware principles. Pilot projects, like those at Maseno and Strathmore University, have highlighted the importance of standardized validation and documentation processes, emphasizing the need for robust governance structures. Moving forward, the DQF will be key in advancing data quality practices across institutions, with continued focus on training, tool deployment, and knowledge sharing.

As data becomes ever more central to decision-making, the DQF provides a strategic foundation for ensuring its integrity, relevance, and impact