Data and Dataset Quality for Artificial Intelligence Systems

Domenico Natale[1]

[1] Chair of Software engineering Italian Technical Commission UNI TC 504, mirror of SC7/WG6; UNI TC 533 AI; member of CEN/CLC JTC 21 AI/WG3 Engineering aspects

Abstract

This paper deals with the standard on data quality for artificial intelligence (AI) with the extension of the concept of data to datasets. It presents the evolution from SQuaRE born in traditional environment towards AI considering the new characteristics of datasets essential for machine learning (ML) and AI applications.

Keywords

Artificial intelligence, data, dataset, quality models, characteristics, measures, evaluation

1. Introduction

The document aims to promote a common language and a culture of data quality. AI systems, in particular machine learning, require data, knowledge and practical experiences to be effective. Systems can be aided by the provision of high-quality data and datasets, including algorithms to process them.

The emergence of AI systems has necessitated a broad reassessment of data quality. This paper discusses the growing importance of data quality when considering AI systems, highlighting the definitions provided in ISO/IEC 5259 series and several data-related standards.

The management involved can adopt quality models for data and provide new and adequate processes, in order to guarantee the expected results.

This paper discusses data quality standards in artificial intelligence (AI) systems. Building on traditional data quality models such as ISO/IEC 25012 and ISO/IEC 25024, it proposes how these standards are evolving to meet also the unique demands of AI and machine learning (ML) applications.

Since 2008, the most complete standard on data quality product is considered the ISO/IEC 25012 "Data quality model" [1], complemented by the standard ISO/IEC 25024 "Measurement of data quality" [2], also important for evaluation. With the advent of Artificial Intelligence, the data quality model has been extended to consider not only individual data, but also groups of data (datasets) necessary for some algorithms, and for machines learning and procedures.

For this reason, ISO/IEC 25012, and ISO/IEC 25024, developed by SC7/WG6, have been imported into ISO/IEC 5259-2 "Data quality measures" [3], with the addition of quality characteristics related to datasets, managed by SC42/WG2. The relationship between the dataset quality model and measures can improve the performance of the AI system and the design of quality characteristics corresponding to the specific AI requirements. The names of some data measures are cited in paragraph 3.2, as well as some results combined with software aspects useful for evaluation of AI systems [4].

2. Evolution

Why do we talking about data quality standards? Because standards play an important role in offering better products, harmonizing terms and processes, facilitating communication, enabling synergies, increasing interoperability of systems and data exchange, with common semantics and with shared metadata, contributing to the reduction of costs and incidents.

It must be considered that data is not a digital artefact, as often happens with software, but above all represents the fact itself and the criterion of truth for each of its aspects. Data is therefore reality, the interpretable representation of what exists. They are the elements on which decisions, improvements and developments can be based. Data, which follows standard rules, are not a convention, but provide information and knowledge of reality. To describe the concept of data quality, standard 25012 defined 15 characteristics, inherent and system dependent. Standard 25024 has defined 63 measures to support the quantification of the level of quality achieved.

Thus, it is happening that AI perspective strengthens data and data quality putting their features at the center of interest, but also requiring the extension of quality and measures to datasets, adding 9 characteristics and some measures.

For AI systems, the qualitative aspects listed below are encounter with new qualitative characteristics: representativeness (with respect to the population to which the AI is intended, refer without discrimination), identifiability or de-identifiability, balance, similarity,

IWESQ'24: International Workshop on Experience with SQuaRE Family and its Future Direction, December 03, 2024, Chongqing, CN

dnatale51@gmail.com (D. Natale)

diversity, effectiveness, relevance, timeliness, auditability, as defined in 5259-2 for machine learning.

The various aspects mentioned highlight the need for specific legislative governance, processes, coordination of products and technical approaches that include training, management preparation, new processes and dissemination of product quality principles for the end user. Data governance will need to ensure, among other things, transparency, stakeholder collaboration and accountability in the different areas of data management.

Progress must be made quickly due to the pervasive impact of AI, considering data and information as an aspect of knowledge, to be explored and recognized from different points of view:

a) *Constraints* : laws, legal and technical quality requirements, human rights, ethics, quality models, methodologies, governance, social impacts, monitoring quality in use;

b) *Development*: management of systems that process input and output data **,** assessment of data and software quality, taking into account cybersecurity, risk classes and automation;

c) *Technologies*: cloud, big data, quantum systems, neural networks, language management, generative models, machine learning , robots, devices, IOT, sensors, hardware, etc..

For each of these points of view, it is reiterated that data, information and knowledge are essential topics to achieve trustworthiness in the new technology.

3. Estending SQuaRE to AI

To better understand the scope of the models developed by SC7/WG6 in ISO/IEC 25000 family [5], widely used in various countries and in Italy, we have to move from SC7 to the context of SC42 which has identified additional quality aspects for datasets. The data document developed by SC42 has not limited itself to the quality of the data described in 5259-2, but also to the quality of other document phases related to the machine learning. Data quality are included in a structured and complete view of the standards:

- 5259-1: overview, terminology, examples
- 5259-2: data quality measures
- 5259-3: data quality management requirements and guidelines
- 5259-4: data quality process framework
- 5259-5: data quality governance framework
- CD TR 5259-6: visualization framework for data quality

This paper summarizes in paragraph 3.2 the essential aspects of the standard 5259-2 and its characteristics.

3.1. Data quality

ISO/IEC 25012 defines a quality model for data, stored in a structured format within a computer system. It defines fifteen quality characteristics for target data used by humans and systems. It considers all data types (e.g. characters strings, texts, dates, numbers, images, sounds, etc.); the scope does not include data produced by embedded devices or real time sensors that are not retained for further processing or historical purposes. The data quality characteristics defined in 25012 are divided in inherent and system dependent. From the inherent point of view, data have the intrinsic potential to satisfy stated and implied needs; from the system dependent point of view, the quality of data depends on the technical domain in which data are used.

The fifteen data quality characteristics defined are classified in the following:

- Inherent: *accuracy, completeness, consistency, credibility, currentness;*
- Inherent and system dependent**:** *accessibility, compliance, confidentiality, efficiency, precision, traceability, understandability;*
- System dependent**:** *availability, portability, recoverability.*

ISO/IEC 25024 specifies 63 measures to quantify the data quality level in relation to characteristics. 5259-2 describes 42 measures imported from 25024, some adopted as is, some adapted for artificial intelligence.

3.2. Datasets quality

The new quality characteristics od datasets defined in 5259-2 can be divided according to the following point of view:
Technical: *Auditability*
Legal: *Identifiability*
Reality oriented: *Balance, Diversity, Effectiveness, Relevance, Representativenes, Similarity, Timeliness*

In addition, the aspects of provenance and randomness are considered in 5259-2 as examples of essential factors for preparation of datasets to reinforce the credibility characteristic.

Regarding the quality characteristics of datasets in 5259-2, measures to quantify quality are also reported. Names of the measures described in the standard for each characteristics are related to: *audited records, resolution balance, balance of images, label proportion balance, category size diversity, category size effectiveness, label effectiveness, identifiability ratio, relevance data in a* *record, representativeness ratio, sample similarity, timeliness of data items.*

Regarding the aspects of balance, representativeness and risks of bias (unwanted discrimination of datasets), other examples are also available from literature [6].

For data quality control, measurement is very important to support quality assessment.
Measurement is a set of operations having the object of determining a value of a measure. The most relevant measures in the contest of data quality are related to the level of quality achieved (threshold) both for individual data and for the group of data in a dataset.
The adoption of the data quality model is essential for verification and evaluation. When following some specific steps to design, control and manage data quality, develop and use AI systems, it is necessary to take into account a detailed data life cycle for different purposes (e.g. data governance, data acquisition, format normalization, data bias, sources integration, storage, maintenance, dismissioning).

Therefore, entire data quality models can be considered an analytical guide for data preparation and not only a useful instrument for evaluation and assessment of a single product.

While data quality concepts and models currently follow a predominantly deterministic orientation in 5259-2, functional measures to quantify datasets quality levels can be extended in the future also to non-deterministic approaches introducing statistical probability methods.

4. Propagation of data quality models

Data quality models can be considered a fundamental topic for AI that links new technology to various existing international standards. For a general operational view we should also consider other standards to be added to those mentioned, such as at least the following, wihich mention SQuaRE standards and data models in the text and mentioned in the bibliography: ISO/IEC 42001:2023 on AI management system, ISO/IEC 25059:2023 on quality model for AI system, ISO/IEC TS 12791:2024 on treatment of unwanted bias, ISO/IEC TR 24027:2021 on bias in AI systems.

ISO/IEC 42001 specifies requirements and provides guidance for establishing, implementing, maintaining and continually improving an AI management system in the context of an organization. This document is intended for use by an organization that provides or uses products or services that use AI systems. The document is intended to help the organization develop, provide or use AI systems responsibly in pursuing its objectives and meeting applicable requirements, stakeholder obligations related and expectations araising from them, paying attention to data quality in the Annex B [7].

ISO/IEC 25059 outlines a quality model for AI systems and is a specific extension from SQuaRE [8]. The characteristics and sub-characteristics detailed in the model provide consistent terminology for specifying, measuring and evaluating AI system quality. They also provide a set of quality characteristics against which the stated quality requirements can be compared for completeness. For example functional adaptability of systems concerns the degree to which an AI system can accurately learn from data or the outcome of previous actions and use that information in future predictions. Furthermore, in the new sub-characteristic, robustness of reliability, is used to describe the ability of a system to maintain its level of functional correctness under any circumstances, including the presence of unseen, distorted, adversarial or invalid data inputs [9].

ISO/IEC TS 12971 describes how to address unwanted bias in AI systems that use machine learning to perform classification and regression tasks. This document provides mitigation techniques that can be applied throughout the life cycle of the AI system in order to treat unwanted bias. This document is applicable to all types and sizes of organizations [10].

ISO/IEC TR 24027 addresses bias in relation to AI systems, particularly with respect to AI-aided decision-making. Measurement techniques and methods for assessing bias are described, with the aim to address and treat bias-related vulnerabilities. All phases of AI system lifecycle are in the scope, including, but not limited to, data collection, training, continual learning, design, testing, evaluation and implementation [11].

5. AI in Europe

A global view of data and datasets quality is described in the technical report CEN CLC TR 18115 “Data Governance and quality for AI within the European context” [12]. The report, developed for information purposes, describes the essential standards for organizations navigating the digital transformation towards AI, considering quality models, best practices, social aspects, ethics, inclusiveness, healthcare. It is addressed to organizations, agencies, enterprises, developers, universities, data scientists, researchers. It cites standards, experiences and best practices, providing an overview on the standards that aim to promote a common language.

The report is consistent with the EU AI Act Regulation[13], that distinguish within harmonized standards, under development, then compliant with the.regulation and prescriptive. All other existing standards, such as those mentioned in this paper, can be considered, now, complimentary and voluntary. By mapping the connections between articles of AI Act and all standards, we obtain a useful structured catalog of "best practices" for the community of standardizers, legal experts. stakeholders, and users. The map, developed in the environment of CEN/CLC JTC21 with the contribution of UNINFO, hosted by the Italian association AI Open Mind, offers a structured view of legistation and technology. It connects concepts with coherent terminology, from a syntactic and semantic point of view, also allowing the possibility to update the SQuaRE quality models [5]. The AI Regulation contains many topics and articles, such as: purpose (art.1), scope (2), definitions (3), education (4), prohibited AI practices (5), risk management (9), data and governance (10), documentation (11), deployers (13), human oversight (14), quality managenment (17), obligation of importers – distributors (23-24), authorities (28), conformity in third countries (39), harmonized standards (40), conformity assessment (43), CE marking (48), AI office (64), testing (76), fundamental rights (77), etc..

6. Conclusion

The title of this paper is just a tip of the iceberg. The topic of data and the whole AI is huge. It is evolving rapidly. The adoption of quality models for data and datasets and datasets and the measurement activities are important for a complete quality assurance avoiding bias and discrimination, supporting useful AI for an efficient decision support system, avoiding bias and discrimination. In addition, it must be considered as a complement to quality of software to manage the data, including features that are also important such as safety and transparency. The propagation of SQuaRE quality models in many standards confirms the value of 25000 family. Likewise, the adoption of a managed quality system in the organization developing AI systems is becoming mandatory, paving the way for conformity assessment and quality certification. It is important to underline the consistency of standards and compliance with the law, improving the efficiency of AI systems production, respecting ethical aspects, human centric vision, inclusiveness and confiming trustworthiness.

References

[1] ISO/IEC 25012:2008 Software engineering - Software product quality Requirements and Evaluation (SQuaRE) - Data quality model

[2] ISO/IEC 25024:2015 Systems and software engineering – Systems and software quality Requirements and Evaluation (SQuaRE) - Measurement of data quality

[3] ISO/IEC 5259-2:2024 Artificial Intelligence - Data quality for analytics and machine learning (ML) – Part 2: Data quality measures

[4] ISO/IEC TS 25058: 2024 Systems and software engineering - Systems and software quality Requirements and Evaluation (SQuaRE) - Guidance for quality evaluation of artificial intelligence (AI) systems

[5] ISO/IEC 25000:2014 Systems and software quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE (summary www.iso25000.it)

[6] A. Simonetta, A. Trenta, M.C. Paoletti, A. Vetrò, Metrics for identifying bias in datasets", CEUR, Workshop Proceedings, 2021

[7] ISO/IEC 42001:2023 Information technology - Artificial intelligence – Management system

[8] D. Natale, Extension of ISO/IEC 25000 Quality Models to the Context of Artificial Intelligence, CEUR-WS, 2022

[9] ISO/IEC 25059:2023 Software engineering - System and software Quality Requirements and Evaluation (SQuaRE) – Quality model for AI systems

[10] ISO/IEC TS 12791:2024 Information technology - Artificial intelligence - Treatment of unwanted bias in classification and regression machine learning tasks

[11] ISO/IEC TR 24027:2021 Information technology - Artificial intelligence (AI) - Bias in AI systems and AI aided decision making

[12] CEN/CLC TR 18115:2024 "Data Governance and quality for AI within the European Context"

[13] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024