

Enterprise Data Trust Score: A Unified AI Framework for Governance, Lineage, Quality, and Compliance

Himanshu Agarwal

Director, Data Quality and Governance.

Abstract

Organizations rely on data to make decisions but many organizations are unable to say how much they should rely on their own data. Quality issues with the data, lack of lineage clarity, poor governance, and regulatory compliance lapses diminish data asset confidence within the enterprise. In this paper, the authors suggest a single artificial intelligence system (Enterprise Data Trust Score) as an integration of governance, lineage, quality and compliance into a single measurable score. The model considers trust as a composite measure constructed out of four weighted aspects. Every part is rated through fixed measures and an artificial intelligence layer updates the measurements depending on the sensitivity and use of the data by the business. The paper has given the framework architecture together with a scoring model which depicts a structure of the framework and the scoring pipeline. The findings indicate that consolidated score provides data stewards with a single figure to measure and enhance. The paper ends by discussing the limitations and future work implications.

Keywords Trust Score, Lineage, Compliance, Blockchain, Governance, Artificial Intelligence.

1. Introduction

Building on large quantities of data is produced and consumed by enterprises on a daily basis. This information aids analytics, machine learning frameworks, monetary reporting, and regulatory reports. The decisions made based on the poor quality of data or lack of clarity regarding the origin of data have implicit risk. It relies on the way the information is controlled, its origin, and its accuracy/completeness as well as whether its application is regulated by the legislation.

This issue is partially investigated by a number of researchers. There are others which are related to blockchain based trust of shared data (Rouhani & Deters, 2021). Other literature dwells on profiling of quality of large sets (Taleb et al., 2021). Other studies concentrate only on the processes of governance and some are only concerned with quality measurement in the domain of automated driving. Lineage and provenance have been explored individually, with blockchain typically serving as tamper proof records (Sigwart et al., 2020). Very little work considers governance, lineage, quality and compliance to be a unitary score that is measurable.

Such a gap poses an effective issue to businesses. There are scores of isolated dimensions, however, they lack a combined and weighted score that can be used to describe the complete state of trust. This is the gap that is dealt with in this paper. There are three contributions of the paper. This paper suggests a model that combines four dimensions of trust, into a single Enterprise Data Trust Score. The article proposes a scoring model three-equation model whereby two weighs the dimension scores with adjustable weights. The results presented in the paper are in table form and a discussion provided on how the score becomes in various conditions of the data.

The remainder of the paper is formatted in the following way. Related work is discussed in section 2. Section 3 refers to the problem and reason why we should have unified score. In section 4, the proposed framework (architecture and the scoring equations) is provided. Section 5 presents results. Section 6 will draw a comparison of the proposed framework and current methods. Limitations and future work are mentioned in Section 7. The paper ends with section 8.

2. Literature Review

Models that aid trust in shared data have been researched, which is based on blockchain. One of them constructs trust based on reputation, endorsement and confidence factors and varies the

amount of people who validate the transaction based on the calculated trust value (Rouhani & Deters, 2021). According to this model, it is demonstrated that there is a possibility of computing trust and that the computation is flexible to risk of the transaction. But the model is primarily concerned with data sharing among parties and is not about an extended governance or compliance with the regulations.

Large datasets quality management has also been addressed. A framework that is created based on a Big Data Quality Profile would store quality rules, attributes and scores and apply profiling prior and after data processing (Taleb et al., 2021). This can enhance the rate of quality checks but treats quality as an individual issue as compared to governance and lineage. The concept of data governance has also been researched independently. One of the frameworks is a zoned data lake that has metadata artifacts that assist automatic governance activities of numerous data sources (Nadal et al., 2022). This reference can help to plan the data pipelines, but it lacks a quantitative measure that indicates trust in dimensions.

Automated decision systems governance is also discussed. An architecture that is constructed around thirteen design principles would encourage data and algorithm stewardship, external inspection under controlled access and data sharing (Janssen et al., 2020). This framework has good principles on governance, but fails to link governance with a numeric score of trust. Quality requirements in safety critical environments have been investigated by way of a design science approach to automated driving systems. That work raises a workflow, a list of challenges, and list of quality attributes all of which have been proven by using focus groups (Pradhan et al., 2023). The methodology is described on the basis of quality management but it does not go further to the lineage or compliance scoring.

Data about provenance and lineage are discussed regarding Internet of Things data. A blockchain smart contracts based, layered architecture documents the provenance of information and its history in a tamper resistant manner (Sigwart et al., 2020). This solution can help address the trust issue regarding provenance records but is restricted to Internet of Things applications and lacks interrelations with governance and compliance regulations utilized by the enterprises.

Compliance research has examined consent with data protection regulation. In one of the studies, it is suggested to create datasets which have already met the recorded consent, with the help of formal representations of consent associated with the purpose of processing (Debruyne et al., 2020). This approach minimises any post facto checking of compliance, but only considers the consent part of compliance, and not the data quality or lineage.

In addition to data owners and users, stakeholder concerns in data sharing have also been examined. A qualitative study of people who post data as crowdfunders shows that these people recognize the worth of the fair use of data, but fear non-payment on time, as well as lack of confidence in the organizations that handle the data (Gomer & Simperl, 2020). The present paper broadens the scope of discussion of the trust of data to the issues of labor concerns, which are missing in technical trust frameworks. Review of research data and the focus have been on quality of the data quality, considering its key features, as well as governance of both corporate and academic environments (Koltay, 2020). This review affirms the fact that quality and governance are closely related to each other and this validates the argument of combining the two with a single score considering them as a whole instead of them being individual.

A data trusts as a service (a cloud-based platform) enables a data subject to view data usage, and to control access (via policies) (Lomotey et al., 2022). This platform employs blockchain to proveance and events to monitor the data streams and reports enhanced transparency in sharing with several parties. This platform is shown to have provenance, access control, and monitoring that can be combined, and suggests that a score structure should be designed as a unit.

In the financial sector, regulatory pressure on quality of data, and their lineage has been directly studied. A regulated financial data platform is analyzed as to how the quality of data can be captured and tracked and lineage artifacts documented to meet the requirements of regulators with regard to accuracy, completeness, consistency and timeliness (Davuluri, 2020). The work is one of the few literatures which relates quality and lineage to compliance obligation which is a great foundation of the compliance element that will be later presented in this paper.

The concept of trust-based governance itself has been doubted as well. A more sceptical examination of trust-based models of data governance questions the notion of accountability, as well as the risks of having a vested concentration of control in one trusted entity (Van Der Sloot &

Keymolen, 2022). The warning of this review is applicable to any framework, even the model presented here, as the false pretence of an effective trust score might be created when this concept is not accompanied by an appropriate supervision.

Spatial data have also been given the same level of data trusts. One of the studies defines five main characteristics needed to have a spatial data trust, and associates these characteristics with the existing literature on spatial data infrastructures and location privacy (Radošević et al., 2023). This paper is concerned with spatial data, but the outlined features like the presence of clean governance and privacy can apply to the overall enterprise data trust design.

A working group of Canadian health data organizations has been able to provide practical advice on how to establish data trusts. This team established twelve data trusts minimums, such as responsible governing body, transparent data management procedures, training and continuing stakeholder involvement (Paprica et al., 2020). The governance component suggested in this paper is similar to the minimum requirements of the framework suggested in this paper.

Data preparation and feature engineering demonstrates the data flow of raw data through the mapping and extraction processes, and quality inspection stages to create a useable feature to be used in machine learning (Johnson et al., 2024). This signifies the importance of the aspect that quality must be measured in various points as opposed to measuring it at one end of consumption. The studies reviewed reflect on effective work at an individual level on the issues of governance, lineage, quality and compliance. All the considered frameworks do not unify all four dimensions into an adjustable score run by artificial intelligence. This is why this gap inspires the framework to be proposed in this paper.

3. Problem Statement and Motivation

There are different teams and different tools which enterprise usually use to govern, lineage, quality, and compliance. A high score of accuracy of a dataset may be reported by a quality team. The same dataset data might be reported by a compliance team that there is an unresolved issue of consent. In the absence of a combination view, decision makers will have no easy way of determining the safety of the dataset to be used.

This division brings about three feasible issues. No single figure reflects the amount of trust in all of the four dimensions so leaders need to read a variety of reports to make a judgement. The emphasis of each dimension is usually predetermined and will not be varied according to the sensitivity and importance of data. A majority of the currently existing scoring methods are non-dynamic and do not incorporate artificial intelligence to change weights and discern trends of hazard in a large volume of sets of information.

Enterprise Data Trust Score framework takes care of these 3 issues. It yields a single score at each dataset. It enables the dimensions weights to be varied with respect to context. It works with a machine learning layer to learn historical results and control further behaviour of scoring.

4. Proposed Enterprise Data Trust Score Framework

4.1 Framework Overview

There are four elements of the framework. These include leadership, family, excellence and conformity. The scoring of each component is in the range of 0-1. A weighting layer of artificial intelligence uses the four pieces of scores as a single Enterprise Data Trust Score that is also zero to one.

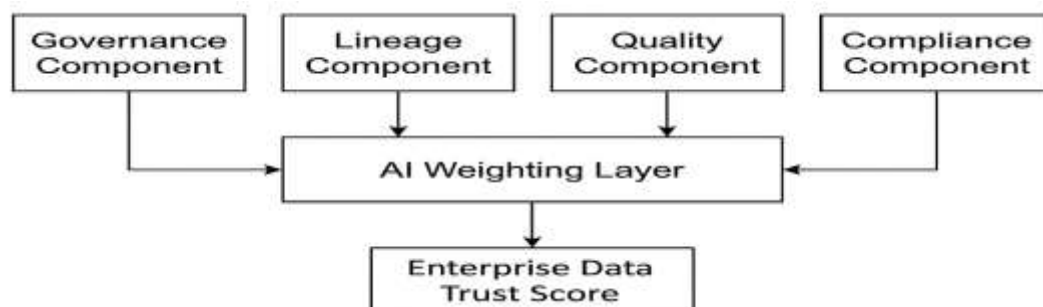


Figure 1 Enterprise Data Trust Score framework

Figure 1 can be interpreted as the flow of information in both ways, that is, starting with the raw data sources and towards the four scoring components into the overall score. This type of structure enables the respective components to be enhanced individually at a time and yet, yield one ultimate outcome.

4.2 Governance Component

Governance component relates to the existence of an accountable owner, a formal policy and a defined process of access control of a dataset. This element is based on the lowest requirements detected to data trusts which affirm that there must be a responsible governing body and transparent data management procedures (Paprica et al., 2020).

The four checks are put together to construct the governance score. These include the assignment of ownership, policy documentation, enforcement of access control and engagement of stakeholders. All the checks will be graded as zero, half or one, according to the absence, partial fulfillment or full fulfillment of the requirement. The score of the four checks are averaged to come up with the governance score.

4.3 Lineage Component

The lineage component is used to gauge the extent to which the history and source of a dataset can be used. This element is inspired by studies of provenance of Internet of Things data based on blockchains that indicate that records of provenance resistant to tampering enhance trust (Sigwart et al., 2020). It is also based on the study of the financial sector that showed that lineage artifacts are a field that requires additional investment although quality monitoring already has reached maturity in several places (Davuluri, 2020).

There are three checks that are used to construct the lineage score. These include the traceability of the sources, logging of its transformation and resilience to tampering of the recorded history. A data set that has complete audit history of their origin to present form has a high lineage score. A data set lacking or having unverifiable history scores low.

4.4 Quality Component

Quality component is the level of accuracy, completeness, consistency and timeliness of the data. This framework is based on the four elements of data quality of financial data governance study and aligns with the quality profiling methodology with large datasets wherein quality regulations produce scores of specified attributes (Taleb et al., 2021; Davuluri, 2020). The quality aspect is also an indication of the design science study on automated driving that demonstrates that quality requirements need to be continuously tested as opposed to being checked only once (Pradhan et al., 2023).

4.5 Compliance Component

The compliance component gauges the compliance of the dataset with the requirements in legislation, as well as regulatory rules, such as consent rules. The work on building datasets that adhere to documented consent at the outset informs this component and part of the approach mitigates the necessity to do compliance checks after the data has already been utilized (Debruyne et al., 2020). It also captures limitations of the trust-based governance which highlight that compliance cannot merely be based on reputation of the holder of the data and that it must be verifiable (Van Der Sloot & Keymolen, 2022).

Evaluations of consent validity, regulatory mapping, retention rules and audit readiness (checks) are constructed to make the compliance score. All checks are zero, half or one with the same scoring method as in the governance element.

4.6 AI Weighting and Score Aggregation Model

After the above computations are made of the four component scores, they should be added to single overall score. The simple average will consider all four dimensions as equally significant to each dataset, which is unrealistic. A dataset to be provided in the financial reporting might require a greater ratio on compliance whereas a dataset to be utilized in internal analytics might require greater ratio on quality.

The framework thus applies adjustable weights which are initially given by data stewards and gradually optimized over time by a machine learning model based on past experiences on

incidents and audit results. The overall Enterprise Data Trust Score can be defined as in Equation 1.

$$DTS = w_G G + w_L L + w_Q Q + w_C C \quad (1)$$

DTS is Enterprise Data Trust Score according to Equation 1. G is the governance score, L is lineage score, Q is quality score, and C is compliance score. The weights of the governance, lineage, quality, and compliance are denoted by w_G , w_L , w_Q and w_C . The weights have to meet the following requirement.

$$w_G + w_L + w_Q + w_C = 1 \quad (2)$$

The quality score is constructed of numerous quality check, as quality is hardly ever a one-dimensional score. The average quality check of n individual quality checks is given in equation 3 as the quality score.

$$Q = \frac{1}{n} \sum_{i=1}^n q_i \quad (3)$$

q_i in Equation 3 will be the value of one quality check out of the possible n quality checks, e.g. accuracy out of completeness and the n is the sum of the number of quality checks applied to the dataset. With this structure, it will be possible to add new quality checks without alteration in the overall method of aggregation.

4.7 Architecture

The structure of the architecture divides the data gathering, scoring and reporting into different layers. Raw data are made available by connectors which are connected to the source systems. The lineage information is stored in a metadata layer as data is transferred between pipelines, and is the same datatype as the metadata artifacts found in automated governance systems (Nadal et al., 2022). A scoring engine realizes the checks, and calculates the component scores. To achieve the final score, Enterprise Data Trust, a weighting engine that is trained based on machine learning is applied with the help of Equation 1. The score is reported to the data stewards and compliance officers to a reporting layer.

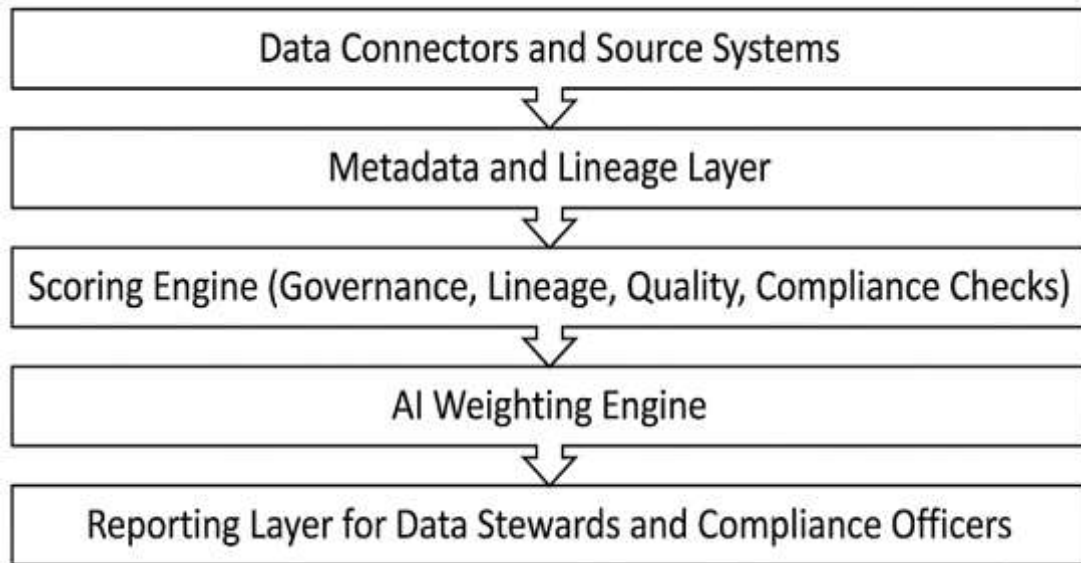


Figure 2 Layered architecture

5. Results

In order to demonstrate the behavior of the framework, the scores of three sample datasets were obtained in accordance with the technique presented in Section 4. Table 1 shows the component scores, and the ultimate Enterprise Data Trust Score of each dataset. The values of the weights in this graph are $w_G = 0.25$, $w_L = 0.20$, $w_Q = 0.30$ and $w_C = 0.25$.

Table 1 Enterprise Data Trust Scores

Dataset	Governance Score	Lineage Score	Quality Score	Compliance Score	Enterprise Data Trust Score
Customer Transactions	0.80	0.70	0.85	0.90	0.815
Marketing Campaign Data	0.60	0.50	0.65	0.55	0.582
Sensor Feed Data	0.40	0.45	0.55	0.35	0.437

The amounts of Table 1 are calculated with the help of equation 1. The Customer Transactions dataset is the highest scoring with it having the highest score in all four dimensions overall. The Sensor Feed Data has the lowest score, mostly due to the poor governance and poor compliance, although its quality score is nearer to the middle range. This tendency is similar to the issue that has emerged in provenance studies where Internet of Things data do not tend to exhibit robust governance mechanisms despite having a satisfactory technical quality of the readings (Sigwart et al., 2020).

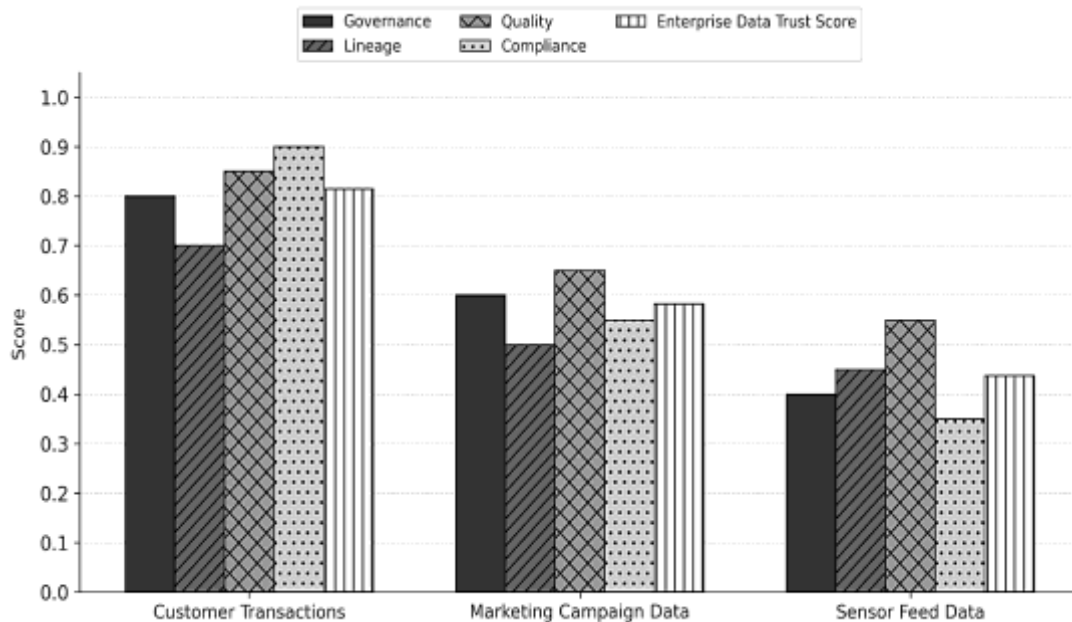


Figure 3 Comparison of components score

This pattern can be visualized in figure 3 enabling data stewards to look and immediately understand which dimension is, on average, keeping the overall score of a particular dataset down.

6. Comparative Analysis

Some of the existing approaches mentioned in the literature review are compared with the framework of the Enterprise Data Trust Score in Table 2 below. Comparison of qualitative ratings has been done as the frameworks being compared are not on a similar numeric scale.

Table 2 Qualitative comparison

Framework	Governance Coverage	Lineage Coverage	Quality Coverage	Compliance Coverage	Combined Score
Blockchain Data Trust Framework	Partial	Strong	Partial	Weak	No

(Rouhani & Deters, 2021)					
Big Data Quality Profile (Taleb et al., 2021)	Weak	Weak	Strong	Weak	No
Automated Data Governance Framework (Nadal et al., 2022)	Strong	Partial	Weak	Weak	No
IoT Provenance Framework (Sigwart et al., 2020)	Weak	Strong	Weak	Weak	No
Consent Compliant Dataset Generation (Debruyne et al., 2020)	Weak	Partial	Weak	Strong	No
Enterprise Data Trust Score Framework (proposed)	Strong	Strong	Strong	Strong	Yes

Table 2 indicates that a majority of the current frameworks used focus on one or two aspects. This is the only proposed framework of the four dimensions that merge the four dimensions into a unified score. It substantiates the main argument of the paper; that is, a single score is necessary to address a gap that dimension specific models have.

7. Limitations and Future Work

The scoring checks should be tested on real datasets across business units in future work, to determine whether or not the scoring checks are consistent. The weighting model in Section 4.6 also requires more outcomes of historical audits to be sure that the machine learning layer can be trained. The weights might have to be configured by data stewards in the initial phase without sufficient historical data.

The framework also proposes that the governance, lineage, quality and compliance may be all transformed into one score out of zero to one. Other qualitative elements like trust between organizations elaborated in spatial data trust studies cannot be readily summarized on a single number (Radošević et al., 2023). Further research in work should determine whether or not extra qualitative indicators should be placed near the numeric score instead of being incorporated within the score.

8. Conclusion

This paper has introduced Enterprise Data Trust Score which is a single framework that sums up the fields of governance, lineage, quality and compliance in a single artificial intelligence-based score. The framework is based on solid previous work in each dimension separately such as blockchain based models of trust, quality profiling models, automated governance models, quality management with a focus on safety, provenance models, and compliance methods driven by consent. It also relies on general research concerning stakeholder issues, the quality of research data, and multi-party data sharing systems, data governed by financial regulation, criticisms of trust-based models, the design of data trusts, as well as the minimum data trust requirements and preparation data to train machine learning.

The scoring model proposed provides the enterprises with a single number to track as opposed to fragmented and reports. These findings and the comparison indicate that the framework provides a solution to gaps that are created by single dimension approaches that are there already.

As more tests on actual enterprise data are done, the Enterprise Data Trust Score will be able to aid in much more predictable and transparent decisions regarding what data to trust and utilize.

References

1. Davuluri, P. S. L. N. (2020). Improving data quality and lineage in regulated financial data platforms. *Universal Journal of Finance and Economics*, 1(1), 1–14. <https://doi.org/10.31586/ujfe.2020.1366>
2. Debruyne, C., Pandit, H. J., Lewis, D., & O'Sullivan, D. (2020). "Just-in-time" generation of datasets by considering structured representations of given consent for GDPR compliance. *Knowledge and Information Systems*, 62(9), 3615–3640. <https://doi.org/10.1007/s10115-020-01468-x>
3. Gomer, R. C., & Simperl, E. (2020). Trusts, co-ops, and crowd workers: Could we include crowd data workers as stakeholders in data trust design? *Data & Policy*, 2. <https://doi.org/10.1017/dap.2020.21>
4. Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. *Government Information Quarterly*, 37(3), 101493. <https://doi.org/10.1016/j.giq.2020.101493>
5. Johnson, S. G., Simon, G., & Aliferis, C. (2024). Data preparation, transforms, quality, and management. In *Health informatics* (pp. 377–413). https://doi.org/10.1007/978-3-031-39355-6_8
6. Koltay, T. (2020). Quality of open research data: values, convergences and governance. *Information*, 11(4), 175. <https://doi.org/10.3390/info11040175>
7. Lomotey, R. K., Kumi, S., & Deters, R. (2022). Data Trusts as a Service: Providing a platform for multi-party data sharing. *International Journal of Information Management Data Insights*, 2(1), 100075. <https://doi.org/10.1016/j.jjimei.2022.100075>
8. Nadal, S., Jovanovic, P., Bilalli, B., & Romero, O. (2022). Operationalizing and automating Data Governance. *Journal of Big Data*, 9(1), 117. <https://doi.org/10.1186/s40537-022-00673-5>
9. Paprica, P. A., Sutherland, E., Smith, A., Brudno, M., Cartagena, R. G., Crichlow, M., Courtney, B., Loken, C., McGrail, K. M., Ryan, A., Schull, M. J., Thorogood, A., Virtanen, C., & Yang, K. (2020). Essential requirements for establishing and operating data trusts. *International Journal for Population Data Science*, 5(1), 1353. <https://doi.org/10.23889/ijpds.v5i1.1353>
10. Pradhan, S. K., Heyn, H., & Knauss, E. (2024). Identifying and managing data quality requirements: a design science study in the field of automated driving. *Software Quality Journal*, 32(2), 313–360. <https://doi.org/10.1007/s11219-023-09622-8>
11. Radosevic, N., Duckham, M., Rahaman, M. S., Ho, S., Williams, K., Hashem, T., & Tao, Y. (2023). Spatial data trusts: an emerging governance framework for sharing spatial data. *International Journal of Digital Earth*, 16(1), 1607–1639. <https://doi.org/10.1080/17538947.2023.2200042>
12. Rouhani, S., & Deters, R. (2021). Data Trust Framework using blockchain technology and adaptive transaction validation. *IEEE Access*, 9, 90379–90391. <https://doi.org/10.1109/access.2021.3091327>
13. Sigwart, M., Borkowski, M., Peise, M., Schulte, S., & Tai, S. (2024). A secure and extensible blockchain-based data provenance framework for the Internet of Things. *Personal and Ubiquitous Computing*, 28(1), 309–323. <https://doi.org/10.1007/s00779-020-01417-z>
14. Taleb, I., Serhani, M. A., Bouhaddioui, C., & Dssouli, R. (2021). Big data quality framework: a holistic approach to continuous quality management. *Journal of Big Data*, 8(1). <https://doi.org/10.1186/s40537-021-00468-0>
15. Van Der Sloot, B., & Keymolen, E. (2022). Can we trust trust-based data governance models? *Data & Policy*, 4. <https://doi.org/10.1017/dap.2022.36>