ISSN 2281-5147

EXTRACTING DATA VALUE THROUGH DATA GOVERNANCE

Armend Duzha[*]

Abstract

Harvesting value from data requires an organization-wide approach. Data governance plays an essential role in a heterogenous environment with multiple entities and complex digital infrastructures, enabling organisations to gain a competitive advantage. This research examines a new approach for data governance developed to extract data value respecting the ever-delicate balance between transparency and privacy. In addition, it provides an overview of the key innovations brought in by novel technologies such as Artificial Intelligence, Federated Learning and Blockchain, and how these can be integrated in a data governance program.

Table of Contents

[*] Armend is a Marie Skłodowska-Curie fellow working as an Early Stage Researcher at University of Piraeus, Greece within the Legality Attentive Data Scientists (LeADS) project funded under the EU's Horizon 2020 Research and Innovation Framework) under Grant Agreement no. 956562.
aduzha@unipi.gr

ISSN 2281-5147

Keywords

Data value – Data governance – Federated learning – Blockchain – Internet of Things

1. Introduction

Many organizations consider data as one of the most important assets (Kitchin, 2021). By leveraging data processing, they transform data into valuable information and meaningful insight. This can involve performing calculations, applying statistical analyses, or using machine learning (ML) algorithms. The use of internet-connected smart devices, the so-called Internet of Things (IoT), and the adoption of Artificial Intelligence (AI) in social life and daily activities, have significantly enhanced the need for data and the general complexity of digital systems. Personal data is collected from various sources (see Figure 1) such as sensors, mobile applications, social networks, and digital footprint left by online activities, in continuous and extensive ways, resulting in a vast data flow for any product and service in use. This has created a new wave of applications that allows organisations to offer user-centric services in many different sectors such as smart city and mobility, healthcare and well-being, smart manufacturing, and finance. For example, consumers can use IoT to monitor their home security and overall health parameters, while businesses can monitor in real-time their supply chain, track energy spending, and engage in predictive maintenance of their machines.

ISSN 2281-5147

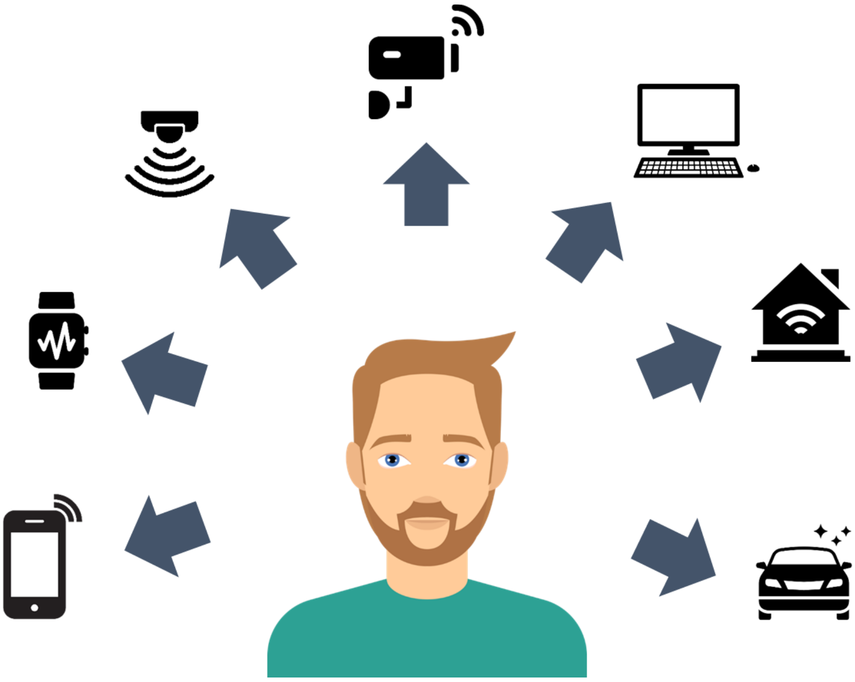

Figure 1: An abstract representation of the Internet of Things

In such a heterogenous environment, data governance plays a crucial role in defining, implementing, and monitoring the context, responsibilities, tools and stakeholders involved throughout the data lifecycle. Therefore, establishing policies, processes and procedures around data and subsequently enacting those to compile and use such data for effective management and decision-making is extremely important. Data governance not only enhances existing products and services but also supports appropriate adjustments during the design and development of new ones.

2. What is Data Governance and why is it important?

The increasing popularity of data governance is closely related with the growing recognition of data value (Zygmuntowski et al, 2021). Before the emergence of AI, data governance primarily focused on the control and management and it constituted a task that was performed mostly by private and large companies. Today, the concept has evolved to encompass "authority" and "control" over the entire data lifecycle with the objective of increasing the data value while minimizing associated risks and costs. In this context, data governance refers to the system of decisions and accountabilities

ISSN 2281-5147

that regulate and guide information-related processes, ensuring adherence to pre-defined policies (Janssen et al, 2020). It outlines permissible activities, assign responsibilities to different entities and actors throughout the value chain, and determine the data to be used, when it can be used, by whom, and for what purposes (Nielsen, 2017).

In the context of AI-based systems, data governance takes a broader role as the system of processes and infrastructures that enable organizations to align AI-enabled technologies with their strategies, objectives, and business values while maximizing the value of data (Mantymaki et al, 2022). Some studies propose the concept of "data governance by design", which facilitate the design of effective data governance frameworks for organizations (Khatri & Brown, 2010).

The emergence of distributed systems, where multiple infrastructures and systems are interconnected, has led to collaborative data processing that involves multiple entities and actors situated in different organisations (Domingue et al, 2019). In this context, decentralised data governance is defined as a set of policies, procedures, and principles that govern the data flows processed by various entities. The decentralised data governance represents a community-based approach for storing, managing, and sharing data, in contrast to a centralized one, where a single entity governs the decision-making throughout the data lifecycle (Greer et al, 2022). Moreover, decentralised data governance ensures compliance with legal, ethical, regulatory, and data protection requirements, specifically the constrains imposed on data processing activities. However, further research to address the challenges posed by decentralisation and distribution are necessary to guarantee that transparency and privacy are not compromised.

3. The Role of Artificial Intelligence

Artificial Intelligence or commonly referred to as AI is one of the dominating trends that affects most industries today, and its impact on data governance is profound. The AI's ability to analyse large datasets fast and efficiently in real-time enables organizations to streamline their data governance practices. Hence, AI solutions can accomplish several tasks such as identification, clustering, classification, and tagging of data, significantly decreasing manual operations.

ISSN 2281-5147

AI-based solutions are adaptable to new data types and sources without requiring extensive re-configuration. Such flexibility is essential in an evolving landscape where both data formats and regulatory requirements keep changing. AI deepens data processing operations by automating existing workflows, making it an indispensable resource in the quest for flexible and scalable data governance.

AI integration into data processing has progressed data governance considerably, thus elevating these tools into intelligent systems capable of autonomous analysis, learning, prediction, and action. With the use of AI, organizations can overcome the complexities of existing digital systems, ensuring their data governance strategies remain effective and responsive to the needs at hand.

As AI systems become increasingly sophisticated, so do the risks associated with data privacy. The same capabilities that make AI powerful also pose significant threats to privacy preservation. To address these challenges, emerging technologies such as Federated Learning and Blockchain are being adopted by the industry.

4. The Power of Federated Learning

Federated Learning (FL) is a recently developed approach for collaborative data processing that is secure and private (Abreha et al, 2022). Let's consider an example from real world to explain how does it work. Google uses FL to build better models for next-word prediction and voice recognition. The company uses the pool of devices (e.g., phones, tables, PCs, watches, and speakers) where Google Assistant is being used. In such a decentralized environment (see Figure 2), user data is stored locally, averting sensitive information from being disclosed with other entities. Instead, each user trains on-device an instance of the ML model, and submits the differences in the parameters to the central server, once the training is completed. The various updates from all the users are then aggregated at the central server, which in return produces an updated global model and distributes it to all users. This iterative process continues until the global model reaches an acceptable level of accuracy or satisfies other criteria defined by Google.

ISSN 2281-5147

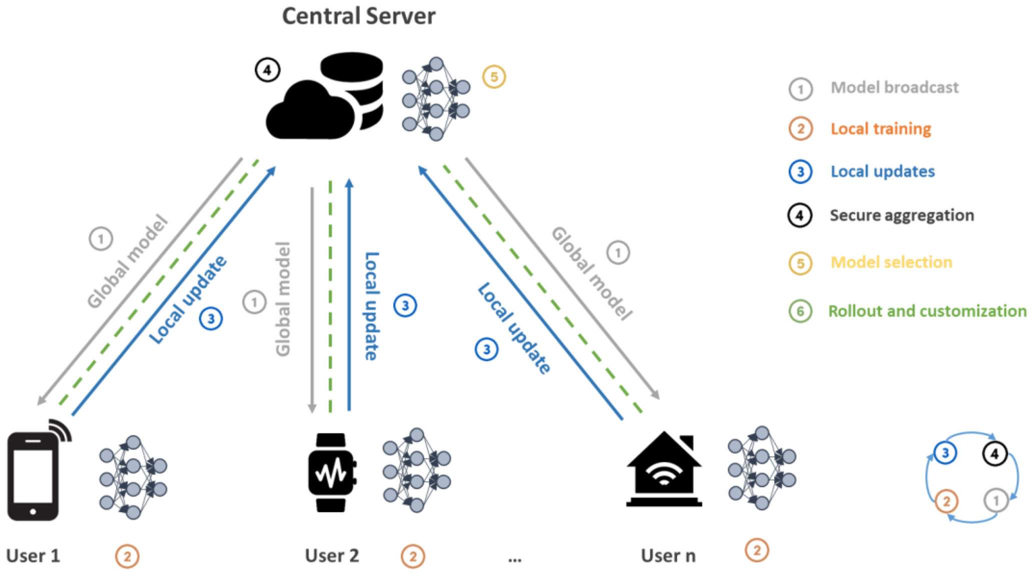

Figure 2: Federated Learning

This inherent structure gives FL the capability to bring together multiple entities in the data processing without the need to disclose any raw data. FL plays a pivotal role in building trust between the involved entities since they can verify at any time that no personal data is exposed, which is crucial for preserving their privacy (Foy et al, 2022).

In terms of resilience, FL enhances the robustness of the system in use in many ways. For instance, it ensures efficient use of data even when network connectivity is poor or one of the users is offline (Bonawitz et al, 2019). Most of the computations happens locally, requiring only intermittent network access to send aggregated model updates. Additionally, FL builds resilience against device failures and data corruption. In our example, if one or more devices goes offline or have corrupted data, the FL process will continue with minimal disruption because it depends on many other devices that continue to work normally, performing local computations. This redundancy makes FL models extremely reliable, ensuring their functionality in any adverse circumstances (Alsamhi et al, 2024).

FL also promotes inclusivity in data processing. By enabling data to remain decentralized, FL allows many data sources as contributors to AI models training without compromising privacy. This can lead to more efficient and potentially more generalized models, since they are trained on a wider variety of data that could not be possible in a centralized setting. Additionally, FL can abide to regulatory requirements, such as data minimisation and storage limitation principles of the GDPR, by limiting sharing of personal data, making it suitable for industries like healthcare, finance, and telecommunications where data privacy is key.

In summary, FL is a powerful methodology for collaborative data processing that enhances privacy, reliability, and transparency. It guarantees that sensitive data are not accessed by third parties, fosters trust among participants, and enables efficient data usage even with limited connectivity. This reliability coupled with FL models' resilience against device failures and data corruption makes it a compelling choice for modern applications.

5. Blockchain: The Backbone of Trust and Transparency

Blockchain, a revolutionary concept that is broadly connected to the use of digital cryptocurrencies like Bitcoin, has turned out to be a powerful tool across different industries ranging from finance to healthcare and beyond (Pilkington, 2016). At its core, it is an immutable distributed ledger that enables data to be secure and tamper-proof, thus forming the core foundation upon which trust can be established.

Its most significant value lies in the ability to produce an irreversible documentation of all transactions and events. Blockchain ensures that once the information is recorded, it cannot be altered or deleted (Zwitter & Hazenberg, 2020). This immutability has profound implications for transparency and trust, since it ensures higher security in data exchanges. Moreover, the very nature of decentralization means that no single entity controls the entire network, instead distributes power across all participants within the blockchain ecosystem, which adds to collaboration and reduce the risk of data misuse.

One of the key benefits of such a decentralised system is that end-users – especially consumers, but also companies – would have much more transparency and control over how their data is used, reclaiming power from big tech and pharma companies that centralise large datasets for competitive benefit.

Healthcare systems in every country and region are struggling with the problem of data siloes, meaning that patients and healthcare providers have an incomplete view of medical histories. In 2016, Johns Hopkins University published research showing that the third leading cause of death in the US was medical errors resulting from poorly coordinated care, such as planned actions not completed as intended or errors of omission in patient records.

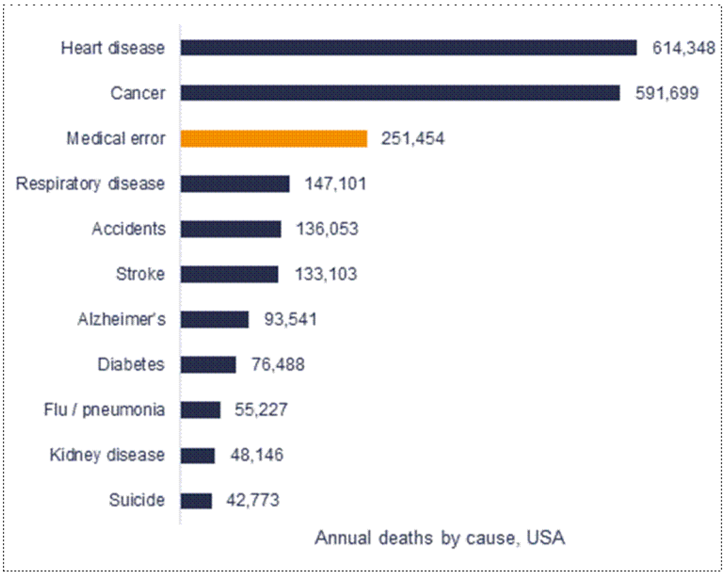

Figure 3: John Hopkins research on medical errors as a proportion of annual deaths in the US, 2016

One potential solution to this problem is creating a blockchain-based system for electronic medical records (EMRs) that can be linked into existing electronic medical record software and act as an overarching, single view of a patient's records. It is crucial to emphasize that actual patient data does not go on the blockchain, but that each new record appended to the blockchain, whether a physician's note, a prescription or a lab result, is translated into a unique hash function – a small string of letters and numbers. Every hash function is unique, and can only be decoded if the person who owns the data – in this case, the patient – gives their consent. In this scenario, every time there is an amendment to a patient medical record, and every time the patient consents to share part of their medical record, it is logged on the blockchain as a unique transaction.

Therefore, in embracing blockchain, organizations are stepping forward towards creating a more trustworthy digital environment. Altogether, blockchain offers unprecedent opportunities to enhance data quality in regards to its integrity, security,

ISSN 2281-5147

and transparency. Due to the distribution of power and record-keeping permanency, it encourages trust among participants in any given sector.

6. Ethical, Legal and Regulatory guidelines

The Ethical, Legal and Regulatory Framework (ELRF) refers to the set of comprehensive rules and practices that should be applied while using AI within organizations. This framework aims to ensure that cutting edge technologies do not violate fundamental rights and values, do not impose bias or discrimination, and are compliant with existing laws, regulations, and ethical standards.

The existing regulatory measures of the European Union include a large number of regulations that tend to facilitate data access and re-use and ensure responsible AI practices. Among others, key regulations include, the *Data Act (DA)* that aims to enhance data re-use by strengthening data sharing and interoperability, the *Data Governance Act (DGA)* that establishes data exchange mechanisms for safe and efficient sharing of data across sectors, and the *AI Act (AIA)*, the European legal framework for AI, which draws standards for just and fair deployment of AI systems. These initiatives are intended to release data currently confined within silos (Patel, 2019) and effectively enforce the data protection framework outlined in the *General Data Protection Regulation (GDPR).*

More specifically, the AIA seeks to play a fundamental role in the fight against inequality, unfair treatment and infringement of right to privacy by AI systems. By safeguarding individual rights, it aims to create the much-needed trust in these technologies, which is a prerequisite towards its widespread acceptance. The AIA thus introduces a risk-based approach categorizing AI systems into different levels of risk:

- *Minimal risk*: systems that pose low or no risk to rights and safety,
- *Limited risk*: systems requiring transparency obligations,
- *High risk*: systems that significantly impact individual rights and safety, subject to stringent regulations,
- *Unacceptable risk*: systems that are prohibited due to their potential of causing harm that is so severe that is unacceptable.

To this end, organizations are required to adapt their strategies to navigate the risk-based framework, ensuring that adopted measures align with the level of potential harm associated with specific AI applications.

The ELRF provides a crucial foundation for the responsible use of AI and Big Data. It ensures that technological advancements initiated by organizations are aligned with societal values and legal requirements. This not only safeguards individual rights but also fosters confidence in AI technologies, paving the way for their broader acceptance and integration into various sectors.

7. The Decentralised Data Governance

Figure 4 presents the high-level architecture of the Decentralised Data Governance framework, which is an evolution of the architecture proposed in the first submission made to AIAI 2022 conference[1]. It has been subsequently improved by integrating the Federated Learning as core component in the data processing layer. Moving outward, it stresses the importance of security, privacy and access control on one side, and ethical, legal and regulatory compliance on the other. These are combined with a blockchain-enabled transactions tracking mechanism, ensuring transparency throughout the system.

[1] Armend Duzha and Dimosthenis Kyriazis, "A Novel Approach for Data Processing and Management in Edge Computing", in 18th International Conference on Artificial Intelligence Applications and Innovations (AIAI 2022), Crete, Greece, June 2022.

ISSN 2281-5147

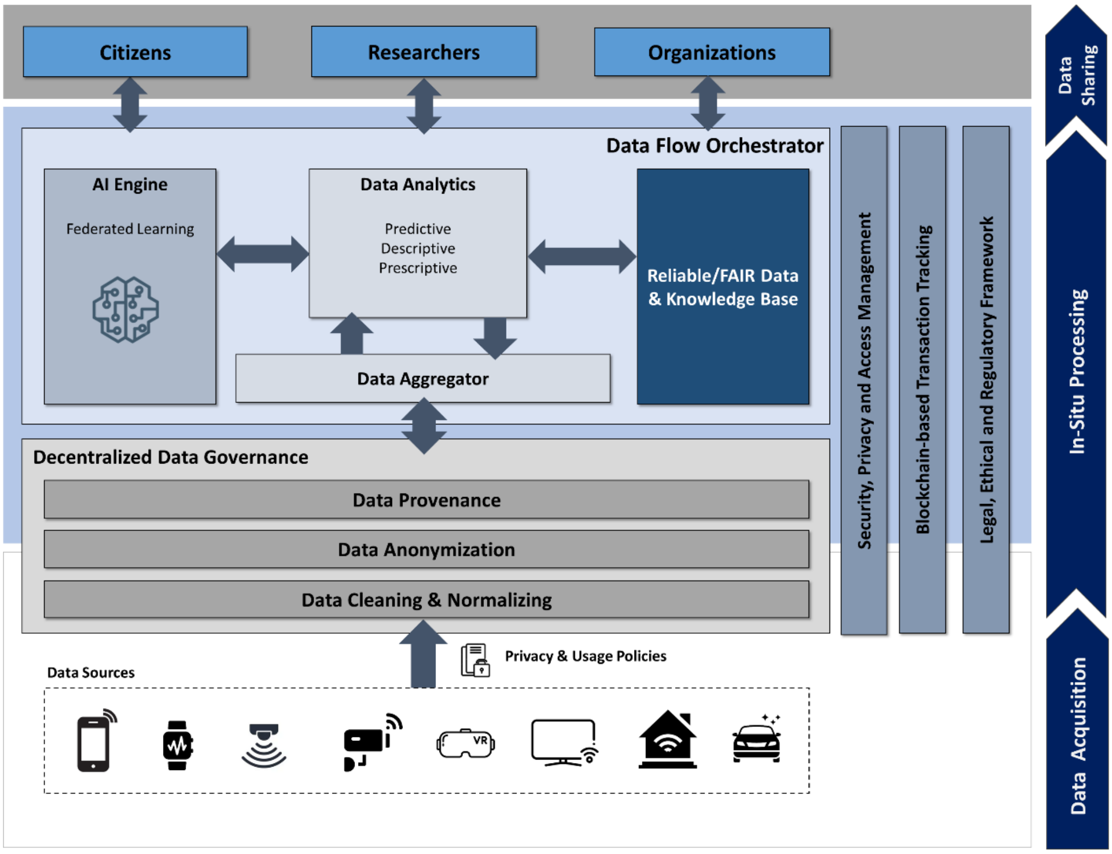

Figure 4: Decentralised Data Governance Framework

The proposed framework consists of three layers: the data layers, the processing layer, and the user layer. The *data layer* is responsible for the preparation of the data to be consumed by the processing layer. This includes cleaning and normalisation so that data is of good quality, as well as an anonymization process before any data activities commences. It follows a decentralised storage approach where each resource remains autonomous. The *processing layer* consists of the AI models and analytics pipelines that process the data derived from the data layer to deliver valuable insights that support informed decisions. Finally, the *user layer* allows to share the processed results and new AI models parameters with data consumers (citizens, researchers and organisations).

The developed framework follows a user-centric approach, empowering and supporting data consumers, be they citizens, researchers, or organisations. It is flexible and adaptable to the needs of the various sectors and aims to maximise the value of data, while minimizing the cost and risks associated with data processing.

ISSN 2281-5147

7.1 Implementation challenges

When initiating a data governance program, organizations may face several challenges that need to be addressed for its successful implementation. Here are some key challenges:

- **Change management and adoption**: persuading business stakeholders to define a data governance program can be a significant obstacle. It requires organisational change efforts that usually involves training, education, and promoting a cultural shift towards data governance practices. Effective communication strategies and stakeholder engagement are crucial for driving adoption and achieving the desired outcomes.
- **Financing**: securing necessary funds can be problematic as it may require determining funding levels for tools to be adopted as part of the program, addressing resource limitations, and understanding how to deliver tangible value. While traditional costs associated with data, such as storage, are relatively quantifiable, assessing its value to the organisation is more complex. The implementation of a decentralised data governance approaches often come with substantial costs and present complex integration and operational activities (Petzold et al, 2020) 0. Moreover, its financing could often be affected by annual variations, making it more challenging to adequately plan it in advance over time.
- **Resource and time constraints**: Data governance activities, such as data collection and definition of data assets, are resource and time-consuming, despite the ability of AI tools to learn data structure and format directly from database structures. Moreover, many organizations nowadays operate in hybrid environments, incorporating both on-premises and cloud-based infrastructures. The distributed data governance requires solutions across multiple platforms and infrastructures, providing consistent governance practices, data policies, and compliance measures.
- **Learning curve, skills and commitment**: Acquiring the necessary skills and the learning curve associated with data governance practices pose time investments and may hinder business adoption, presenting serious commitment challenges.

ISSN 2281-5147

By overcoming these challenges, organizations can unlock the full potential of data governance and realize its benefits in terms of improved data management, enhanced decision-making, and compliance with regulatory requirements.

7.2 Benefits for organisations and individuals

The decentralised data governance can be adopted by several organisation, and more specifically: (i) *private organisations* for planning, designing and implementing data-driven solutions to align their vision and business objectives with their customers' needs; (ii) *public organizations* for organizing, planning, and monitoring the timely provision of appropriate and enhanced policies leading to efficient decision-making and services provided to the citizens; and (iii) *researchers* involved in data processing activities, to facilitate data discovery, interoperability and re-use.

Moreover, data governance can boost the ability of public and private organisations to exploit and monetize their data assets (Ofulue and Benyoucef, 2022), while at the same time developing novel services or enhancing the existing ones.

Finally, data governance changes the approach of assessing the cost related to data, shifting the focus from "How much does processing data cost?" to "How much should our organization invest in data governance?". By adopting a distributed data governance approach, organisations can distribute the implementation risk across the whole data lifecycle. Unlike in-house data processing, distributed data governance offers a flexible support structure that can be adopted to the project's needs, timelines, and requirements.

8. Conclusions and future work

The decentralised data governance represents a necessary step towards responsible and ethical harnessing the value of data. The proposed approach underscores the importance of clear principles, policies, and robust privacy-preservation techniques such as Federated Learning and Blockchain, putting control over how personal data is used into the hands of individuals and organisations. Key legal and normative considerations, such as adherence to the AI Act (AIA) and the General Data Protection Regulation (GDPR), will ensure that data processing within the data lifecycle is conducted legally and ethically.

Looking ahead, the prospects that data governance can truly revolutionize the data economy is great. By facilitating secure and transparent data processing, decentralised data governance can drive innovations in different sectors, including healthcare, smart city and mobility, smart manufacturing, and finance. Its potential is huge, although some challenges still remain – respecting the ever-delicate balance between transparency and privacy and making sure that data processing does not jeopardise individual rights. The framework must continuously evolve in view of such problems, adapting to new types and sources of data, emerging technologies, or regulatory changes.

By raising awareness of data value and governance principles, the benefits of the data governance can be realized while safeguarding privacy and data protection. This decentralised approach will help build trust and support for data-driven innovations, ultimately fostering a more responsible and secure data ecosystem.

9. Selected Readings

R. Kitchin, "The Data Revolution: A Critical Analysis of Big Data, Open Data and Data Infrastructures", SAGE Publications, 2021.

J. J. Zygmuntowski, L. Zoboli, and P. F. Nemitz, "Embedding European values in data governance: a case for public data commons", Internet Policy Review, Vol. 10, No. 3, September 2021. DOI: 10.14763/2021.3.1572

M. Janssen et al. "Data governance: Organizing data for trustworthy Artificial Intelligence", Government Information Quarterly, Vol. 37, No. 3, pp. 101493, July 2020. DOI: 10.1016/j.giq.2020.101493.

O. B. Nielsen, "A comprehensive review of Data Governance literature", Selected Papers of the IRIS, Vol. 3, No. 8, 2017. Available at: https://aisel.aisnet.org/iris2017/3.

M. Mantymaki et al. "Defining organizational AI governance", AI and Ethics, February 2022. DOI: 10.1007/s43681-022-00143-x.

V. Khatri and C. V. Brown, "Designing Data Governance", Communications, Vol. 53, No. 1, pp. 148–152, January 2010.

J. Domingue, A. Third, and M. Ramachandran, "The FARI-TRADE Framework for Assessing Decentralized Data Solutions," in Companion Proceedings of the 2019 World Wide Web Conference (WWW '19), ACM, New York, NY, USA 2019.

S. L. Greer et al., "Centralizing and Decentralizing Governance in the COVID-19 Pandemic: The Politics of Credit and Blame", Health Policy, Vol. 126, No. 5, May 2022.

H. G. Abreha, M. Hayajneh, and M. A. Serhani, "Federated Learning in Edge Computing: A systematic survey," Sensors, Vol. 22, No. 2, January 2022.

M. Foy, D. Martyn, D. Daly, A. Byrne, C. Aguneche, and R. Brennan, "Blockchain-based governance models for COVID-19 digital health certificates: A legal, technical, ethical and security requirements analysis", Procedia Computer Science, No. 198, pp. 662–669, 2022.

S. H. Alsamhi et al., "Federated Learning meets Blockchain in decentralized data sharing: healthcare use case", IEEE Internet of Things Journal, Vol. 11, No. 11, pp. 19602-19615, April 2024.

Marc Pilkington, "Blockchain technology: principles and applications", Research Handbook on Digital Transformations, pp. 225–253. Edward Elgar Publishing, 2016.

M. Foy, D. Martyn, D. Daly, A. Byrne, C. Aguneche, and R. Brennan, "Blockchain-based governance models for COVID-19 digital health certificates: A legal, technical, ethical and security requirements analysis", Procedia Computer Science, No. 198, pp. 662–669, 2022.

A. Zwitter and J. Hazenberg, "Decentralized network governance: Blockchain technology and the future of regulation", Frontiers in Blockchain, Vol. 3, 2020. DOI: 10.3389/fbloc.2020.00012.

R. Latif, M. U. Ahmed, Sh. Tahir, S. Latif, W. Iqbal and A. Ahmad, "A novel trust management model for edge computing", Complex & Intelligent Systems, Vol. 8, pp. 3747–3763, 2022. DOI: 10.1007/s40747-021-00518-3.

B. Petzold, M. Roggendorf, K. Rowshankish, and C. Sporleder, "Designing data governance that delivers value", McKinsey Digital, June 2020.

J. Ofulue and M. Benyoucef, “Data monetization: insights from a technology-enabled literature review and research agenda, Management Review Quarterly, 2022.

ISSN 2281-5147