Data Governance as the Cornerstone of Good Governance in Mega Projects

Andy WL Chung, Smart City Maker Ltd., Hong Kong SAR, China. Email: ac@smartcitymaker.com
Wai Ming To, Macao Polytechnic University, Macao SAR, China. Email: wmto@mpu.edu.mo
Lavin YM Yeung, ESG Matters, Hong Kong SAR, China. Email: lavin@esgmatters.net

Abstract

This paper explores the critical role of data governance in supporting good governance in mega projects, using the conceptual framework of a mega-infrastructure project. The research highlights the importance of establishing a robust data governance plan as a prerequisite for developing an effective and reliable data system. The paper delves into the three pillars of data governance—data quality management, data privacy and security, and regulatory compliance—and examines how they contribute to achieving transparency, accountability, and informed decision-making in mega projects. Through the analysis of these pillars, the paper emphasizes the need for a comprehensive and strategic approach to data governance that lays the foundation for a data system aligned with the principles of good governance. A practical application of the principles, ANDANTE, will be presented.

Summary Statement

This paper explores the critical role of data governance in supporting good governance in mega projects, using the conceptual framework of a mega-infrastructure project.

1. Introduction

Data governance is an evolving concept, driven by rapid developments in information and communication technologies. It is especially critical in mega projects where diverse stakeholders play various roles in the collection, storage, control, processing, use, and dissemination of data. Moreover, ensuring data privacy, security, and compliance with regulations is paramount to protect data integrity and to facilitate proper data use and analysis.

Historically, mega project proponents and their consultants would conceive projects, prepare detailed proposals, and develop tender documents with contributions from sub-consultants and specialists. Such teams might devise data management plans to organize and continuously monitor project progress using collected data. Typically, these teams adopted a traditional governance approach, characterized by a hierarchical, top-down perspective (Sweeney, 2022). This approach can convey “value perceptions” of project proponents and their consultants about what data should be captured and how data should be utilized. However, it often overlooks the broader needs for data management and usage from construction and operational perspectives and does not accommodate new methods for collecting, storing, analyzing, using, and disseminating data from other stakeholders’ perspectives. This article offers a brief review of data governance development using a bibliometric approach. It proposes a comprehensive definition of data governance, introduces a robust governance plan, and outlines three key pillars: data quality management, data privacy and security, and regulatory compliance. It concludes with a proposed framework for data governance and the Automatic Noise Data Management E-system (ANDANTE), which integrates web-based noise and weather monitoring with a CCTV network, ending with some final remarks.

2. Data Governance

Understanding an evolving concept such as data governance from a holistic perspective is challenging. Nevertheless, with the advent of bibliometric science—particularly its methodologies, data sources, and tools—researchers can obtain a panoramic view of specific research topics quickly and objectively (Donthu et al., 2021; To, 2022; Yan et al., 2022). Moreover, bibliometric studies can illuminate trends on research topics, identify the most productive authors, institutions, and countries, and highlight the initial publications and their core themes, as well as the evolution of underlying concepts over time (Chung and To, 2023; To and Chung, 2023; To et al., 2023).

On January 3, 2024, a search was conducted using the term "data governance" in "Article Title, Abstract, Keywords" on Scopus—one of the largest academic indexing databases. This search yielded 2,354 documents, including journal articles, reviews, conference papers, book chapters, books, conference reviews, notes, editorials, short surveys, and others. After excluding 50 conference reviews, 24 notes, 9 editorials, 8 errata, 5 short surveys, and 1 letter, 2,257 documents remained. These included 1,140 journal articles, 132 reviews in journals, 756 conference papers, 45 books, and 184 book chapters. The earliest academic article on “data governance” appeared in year 2005 (Trope and Power, 2005). The number of publications increased to 15 in year 2010, 96 in year 2017, and 484 in year 2023, as illustrated in Figure 1(a).

According to Scopus, the most productive author was Rob Brennan from the School of Computer Science

at University College Dublin, with 17 data governance publications. Boris Otto from the Fraunhofer Institute for Software and Systems Engineering ISST in Germany was the second most productive author, with 14 publications. In terms of institutional affiliations, the University of Oxford led with 37 publications, followed by Delft University of Technology with 29, and the University of Toronto with 28. In a geographic distribution, U.S.-affiliated authors produced 466 data governance publications, Chinese-affiliated authors 323, and UK-affiliated authors 320. Figure 1(b) shows that U.S. and UK authors were most prolific before 2019; however, publications from Chinese authors have surged in recent years (To and Yu, 2020).

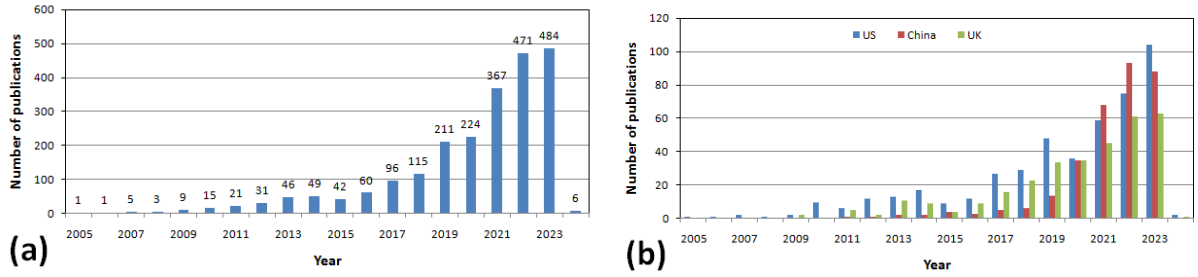

Figure 1. The number of data governance publications from 2005 to 2023 (a) total, and (b) by US, Chinese, and UK authors.

Scopus revealed that the first three articles on data governance, authored by Roland L. Trope, E. Michael Power, and their associates in 2005, 2006, and 2007, approached the subject from a legal, i.e., compliance perspective. Specifically, Trope et al. (2007) highlighted that an organization's data should be actively managed, especially in terms of data security in an increasingly technology-intense environment. Organizations should develop coherent information management strategies, incorporating inputs from business alliances, to meet regulatory requirements. A review of the first ten data governance publications identified by Scopus showed a focus on data quality management, privacy, and information security.

A keyword co-occurrence analysis of the 2,257 selected publications was conducted using VOSviewer (Van Eck and Waltman, 2010). Setting the minimum occurrence of a keyword at 25, 104 out of the 10,718 keywords met this threshold, resulting in three clusters as illustrated in Figure 2. The largest cluster (colored red) included 46 keywords with 'data governance' as the core keyword, encompassing terms like big data, information management, artificial intelligence, decision making, data quality, and data management. The second largest cluster (colored green) included 38 keywords with 'human' as the core keyword, involving terms such as article, adult, data sharing, privacy, ethics, and medical research. The third cluster (colored blue) focused on 'data privacy' and included keywords such as governance, data protection, blockchain, security, and laws and legislation. Notably, the second largest cluster primarily discussed data governance concerning personal medical data.

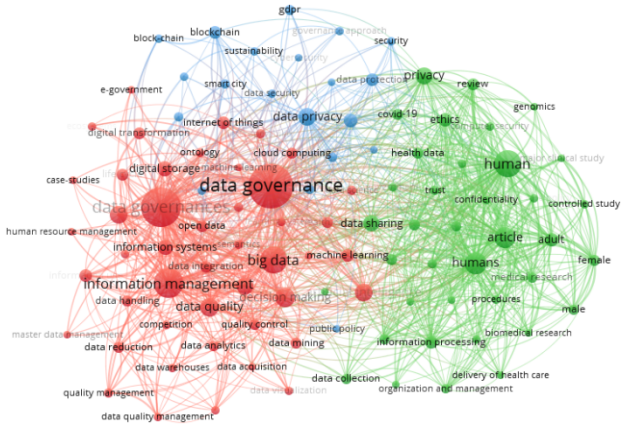

Figure 2. Co-occurrences of keywords using the 2,257 data governance publications.

Through meticulous bibliometric analysis, data governance can be defined as a holistic approach encompassing strategies, policies, structures, processes, procedures, and practices that assist organizations in planning, budgeting, collecting, organizing, storing, retrieving, analyzing, processing, presenting, reporting, and disseminating data and outcomes effectively and efficiently. Moreover, data governance must address data quality management, data privacy and security, and adherence to national and international regulations.

A comprehensive data governance plan involves several steps: (i) assessing data needs from various stakeholders, (ii) defining roles, rights, responsibilities, and obligations for key parties, including different stakeholders involved in data collection, management, control, use, and disposal, and (iii) establishing data policies and standards that dictate how data are to be handled and shared. These policies and standards must comply with regulatory requirements, such as data privacy and protection. Furthermore, a data governance plan should cover its implementation, including the evaluation of different data platforms (and/or databases), applications, analytics, and tools, assessing potential issues and risks. Implementation should be followed by measurement, analysis, and improvement of data governance, using the well-established Deming Plan-Do-Check-Act cycle (Carretero et al., 2016) as depicted in Figure 3.

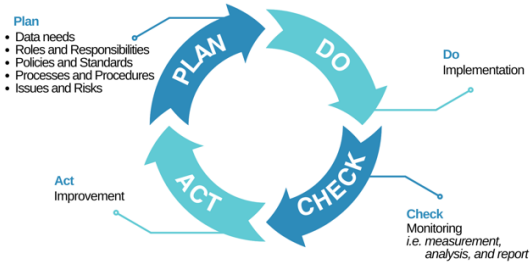

Figure 3. A data governance plan based on Deming Plan-Do-Check-Act cycle.

2.1. Data quality management

As a key pillar of data governance, data quality management ensures that collected data possesses desirable characteristics such as usability and usefulness to data consumers and is free of defects (Fürber, 2016; Redman, 2001). More specifically, data quality measures the extent to which a dataset meets criteria in terms of accuracy, completeness, consistency, uniqueness, validity, and timeliness (IBM, 2023). According to IBM, these dimensions include:

- Accuracy: Characterizes the correctness of data values and the importance of identifying a "source of truth"—acting as a primary data source while other data sources can be used to confirm the agreed value(s).
- Completeness: Reflects the usefulness of the collected data; a high percentage of missing values may lead to misleading or biased analyses.
- Consistency: Evaluates data records from two or more different datasets to ensure that the same or similar conclusions can be obtained without contradictory evidence.
- Uniqueness: Means that duplicate data should be processed and eliminated.
- Validity: Measures the extent to which data meets the required formats established by the organization (or project proponent).
- Timeliness: Characterizes the readiness of data within an expected timeframe. High-quality data is essential for effective evidence-based decision-making to achieve an organization's (or a project's) goals.

Only when high quality of data is available, effective evidence-based decision making can be made in order to achieve an organization's (or a project's) goals.

2.2. Data privacy and security

Data privacy and data security are closely related concepts. Data privacy involves the collection, control, and protection of personal information, focusing on lawful and informed collection, proper handling, storage, processing, usage, and disposal of personal data. Privacy is considered an individual's human right and should not be compromised without knowledge or consent (De Hert and Gutwirth, 2006).

Data security primarily focuses on protecting data from unauthorized third-party access, malicious attacks, and improper exploitation. It encompasses organizational (or project) policies, programs, practices, and processes established to protect data, including personal data. Data security practices may include network security, access control, breach response, encryption, multi-factor authentication, and activity monitoring. In 2019, the International Organization for Standardization (ISO) released ISO/IEC 27701 outlining a privacy information management system for privacy information security (ISO, 2019). This standard assists organizations to establish management systems that support data privacy requirements, particularly for compliance of the European Union General Data Protection Regulation (GDPR).

2.3. Regulatory compliance

Data are important organizational assets. As data often includes personal data, various national, regional, and

international laws govern the collection, storage, management, control, usage, and disposal of personal data. The GDPR, a regional regulation about data protection, covers data privacy and security, imposing obligations on any organization that handles data related to EU citizens or residents (European Union, 2023). Effective since 25 May 2018, the GDPR mandates that organizations consider data protection "by design and by default."

In the US, while there is no federal data privacy law, approximately one-third of the states have begun to pass or enact data privacy legislation. For example, the California Consumer Privacy Act (CCPA) took effect in January 2020, granting individuals the rights to opt-out of data collection, and to access and delete their data, similar to the GDPR (dfinsolutions.com, 2023). Across the Pacific, China's Personal Information Protection Law (PIPL) became effective on 1 November 2021 (Hong Kong's Privacy Commissioner for Personal Data, 2023). Like the GDPR, the PIPL regulates the collection and processing of data from individuals in China, regardless of whether the organizations are based in China.

3. A Proposed Framework and ANDANTE

Figure 4 shows a proposed data governance framework. It demonstrates that data lifecycle begins with a data needs analysis. An organization (or a project proponent) must sit down with its stakeholders to identify what data are required, where and when data are collected, who collect and take charge of the storage, management and control of data, how data are processed and analyzed, how processed data are presented (and disseminated), and how and when data are disposed of (if necessary). Once a data needs analysis is thoroughly conducted with inputs from different stakeholders and a consensus is reached, the organization (or project proponent) shall consider the appointment of people who are responsible for the collection, storage, management and control, processing, analyzing, disseminating, and disposal of data and provide appropriate technical and compliance-related training.

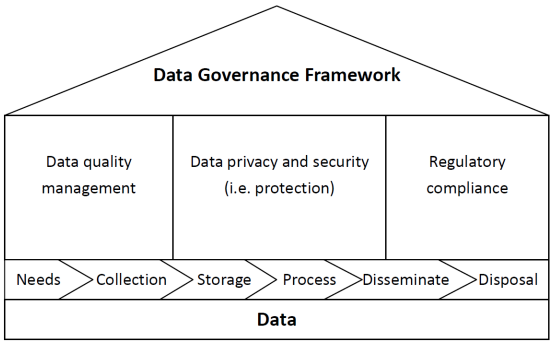

Figure 4. A proposed data governance framework

ANDANTE, a past project in Hong Kong, stands as a testament to the practical application of this framework.

In Hong Kong, mass transportation is critical because it serves about 4 millions passenger journeys a day including millions of residents and travelers (To, 2015). Additionally, mass transit railway is relatively more environmental-friendly than other modes of transport such as public buses and private cars (To et al., 2020). Thus, Hong Kong's mass transit railway has kept expanding over the decades (To, 2015; To et al., 2020). During the construction or extension of rail lines, noise monitoring with timely alerts and the resolution of noise exceedances are crucial for regulatory compliance. An automatic noise data management e-system, namely ANDANTE, has been deployed to meet such requirements (Chung et al., 2012). ANDANTE is a cloud-based online platform that collects noise, visual, and weather data along the construction and extension of rail lines. It consists of noise monitoring stations, CCTV cameras, and weather stations (Chung et al., 2012). Figure 5 shows a screen-shot of ANDANTE and its associated equipment. One of its added features is the enhancement of stakeholders' engagement. ANDANTE enables real-time notifications to all concerned parties when a noise exceedance or noise complaint occurs.

The implementation of ANDANTE's data governance framework not only enhanced operational transparency but also bolstered public trust by providing stakeholders with real-time access to environmental monitoring data.

Figure 5. Screen-shot and associated equipment of ANDANTE

4. Conclusion

Data governance is vital for an organization's survival and prosperity, as it hinges on effective data collection and capturing "the value" of data while accommodating the needs and requirements of various stakeholders. Similarly, the success of a mega project relies on its data management and governance aligning with stakeholder and regulatory expectations. This paper has provided an overview of the evolution of data governance through bibliometric analysis, introduced a robust data governance plan, and described the characteristics of its three main pillars: data quality management, data privacy and security, and regulatory compliance. Additionally, the paper briefly introduces a practical system, ANDANTE, which exemplifies effective data governance.

References

Carretero, A. G., Freitas, A., Cruz-Correia, R. and Piattini, M. (2016). A case study on assessing the organizational maturity of data management, data quality management and data governance by means of MAMD. *Proceedings of ICIQ 2016*, Ciudad Real, Spain, 22-23 June 2016 (pp. 75-84).

Chung, A. W. L. and To, W. M. (2023). A bibliometric study of carbon neutrality: 2001–2022. *HKIE Transactions*, 30(2), 1–11.

Chung, A., Choi, J., Leung, H., Chan, S. and Frommer, G. (2012). ANDANTE – Legal compliance & improving mgt. efficiency. *Proceedings of IAIA 2012 – Energy Future. The Role of Impact Assessment*, Centro de Congressos da Alfandega in Porto, Portugal, 27 May - 1 June 2012.

De Hert, P. and Gutwirth, S. (2006). Privacy, data protection and law enforcement. Opacity of the individual and transparency of power. In Claes, E., Duff, A. and Gutwirth, S. (Ed.). *Privacy and the Criminal Law*. Intersentia, Antwerp/Oxford, pp. 61-104.

Dfinsolutions.com (2023). *Data Protection in Transition: GDPR, CCPA and Comparable Data Protection Laws*. Donnelley Financial Solutions. Available at: https://www.dfinsolutions.com/en-gb/knowledge-hub/article/gdpr-ccpa-and-US-data-privacy-laws?type=pmax&gclid=Cj0KCQiA6vaqBhCbARIsACF9M6mPTfCfJsfMnjoKko-jufgGy2fml6fFre7W4CtekKFP9UVY6UtQr00aAmDIEALw_wcB

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N. and Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133, 285-296.

European Union (2023). *What is GDPR, the EU's New Data Protection Law?* European Union. Available at: https://gdpr.eu/what-is-gdpr/#:~:text=The%20General%20Data%20Protection%20Regulation,to%20people%20in%20the%20EU.

Fürber, C. (2016). Data quality. In *Data Quality Management with Semantic Technologies*. Springer Gabler, Wiesbaden.

Hong Kong's Privacy Commissioner for Personal Data (2023). *Mainland's (China) Personal Information Protection Law*. Privacy Commissioner for Personal Data. Available at: https://www.pcpd.org.hk/english/data_privacy_law/mainland_law/mainland_law.html

IBM (2023). *What is Data Quality?* IBM. Available at: https://www.ibm.com/topics/data-quality#:~:text=the%20next%20step-,What%20is%20data%20quality%3F,governance%20initiatives%20within%20an%20organization.

ISO (2019). *ISO/IEC 27701:2019 Security Techniques – Extension to ISO/IEC 27001 and ISO/IEC 27002 for Privacy Information Management – Requirements and Guidelines*. International Organization for Standardization, Geneva, Switzerland.

Power, E. M. and Trope, R. L. (2006). The 2006 survey of legal developments in data management, privacy, and information security: The continuing evolution of data governance. *Business Lawyer*, 62(1), 251-294.

Redman, T. C. (2001). *Data Quality: The Field Guide*. Digital Press.

Sweeney, K. (2022). *Holistic Data Governance*. Stats NZ. Available at: https://data.govt.nz/assets/Uploads/summary-holistic-data-governance.pdf

To, W. M. (2015). Centrality of an urban rail system. *Urban Rail Transit*, 1(4), 249-256.

To, W. M. (2022). A bibliometric analysis of world issues—Social, political, economic, and environmental dimensions. *World*, 3(3), 619-638.

To, W. M. and Chung, A. W. L. (2023). Carbon-neutrality research in China—Trends and emerging themes. *World*, 4(3), 490-508.

To, W. M. and Yu, B. T. W. (2020). Rise in higher education researchers and academic publications. *Emerald Open Research*, 2, 3. https://doi.org/10.1108/EOR-03-2023-0008

To, W. M., Lee, P. K. C. and Yu, B. T. W. (2020). Sustainability assessment of an urban rail system–The case of Hong Kong. *Journal of Cleaner Production*, 253, 119961.

To, W. M., Yu, B.T. W., Chung, A. W. L. and Chung, D. W. K. (2023). Metaverse: Trend, emerging themes, and future directions. *Transactions of Emerging Telecommunications Technologies*. https://doi.org/10.1002/ett.4912

Trope, R. L. and Power, E. M. (2005). Lessons in data governance: A survey of legal developments in data management, privacy and security. *Business Lawyer*, 61(1), 471-516.

Trope, R. L., Power, E. M., Polley, V. I. and Morley, B. C. (2007). A coherent strategy for data security through data governance. *IEEE Security and Privacy*, 5(3), 32-39.

Van Eck, N. and Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523-538.

Yan, C., Li, H., Pu, R., Deeprasert, J. and Jotikasthira, N. (2022). Knowledge mapping of research data in China: A bibliometric study using visual analysis. *Library Hi Tech*. https://doi.org/10.1108/LHT-11-2020-0285