Unlocking the Business Value of Dark Data: The Importance of Data Governance

Graham Chant[1] and Jennieffer Barr[2]
[1]Stewart Barr and Associates, Adelaide, Australia
[2]Charles Darwin University, Darwin, Australia

graham.chant@sb-a.com.au
jennieffer.barr@cdu.edu.au

Abstract: Hidden amongst data is dark data; unstructured, unanalysed and typically unused data. Dark data is a complex phenomenon as it is typically difficult to access, lacks meaning without human interpretation, yet could add value and benefit to an organisation. Alternatively, a risk of not activating dark data for an organisation could be inaccurate data used for decision making. Dark data provides the opportunity for building knowledge and in turn, can provide a competitive advantage. Whilst awareness of the value of dark data is increasing, accessing and effectively using this data is not straightforward. Dark data engagement brings with it challenges of how to access this data, how to convert this data for it to be usable and who will be accountable to ensure appropriate use of such data. This paper argues that data governance will be necessary to unlock dark data, control the conversion of dark data to usable data and manage appropriate use of such data. This paper also argues that not unlocking dark data may mean business decision-making may be at risk of not having an accurate base to inform decisions. Appropriate use of dark data aims to maximize benefits for a competitive company, but these organizations will need data governance to control and manage how this data is used. This paper will examine data governance to reduce risks associated with dark data, ensure security and integrity of all data, protect data from misuse, and improve dark data processing.

Keywords: Dark data, Data governance, Quality, Security, Unstructured, Big data

1. Introduction

There is almost no limit to the kinds and amount of data that can now be collected (Eryurek et al., 2021). Statista (cited in Duarte 2024) reported that the latest estimates indicate that throughout 2024, approximately 147 zettabytes of data will be generated encompassing newly created, captured, copied, and consumed information. By the end of 2025, the global volume of data is projected to rise further to 181 zettabytes. The challenge for managers and directors of companies will be to manage this amount of data and reduce the risk of increasingly unused or misusing data. Effective data governance will be the key to addressing this challenge.

A variety of data governance definitions have been presented in recent literature. Currently, there is a lack of consensus for use of a consistent definition of data governance. Some use data governance interchangeably with the term information governance (Black et al., 2023). However, according to Bennett (2017) information governance differs from data governance. Bennett (2017) explains that data governance is a subset of information governance and only refers to data. In comparison, information governance includes considering the context of data which then becomes information. This paper will focus on data governance.

Abraham, vom Brocke & Schneider (2019) undertook a literature review of 145 published papers examining definitions. They provided a definition of data governance based on a cross-functional framework for managing data. Like other authors, Abraham, vom Brocke & Schneider's (2019) definition included a variety of aspects of data governance. Defining data governance commonly includes the notion of control of data (Abraham, vom Brocke & Schneider, 2019; Bennett, 2017). Control is asserted through decisions about rights and accountabilities in managing data (Benfeldt, Persson & Madsen, 2019; Bennett, 2017). Control through authority (Abraham, vom Brocke & Schneider, 2019; Bennett, 2017) often referred to as roles and responsibilities (Benfeldt, Persson & Madsen, 2019;) is the process of determining which people in an organisation will have the decision-making power of data governance (Al-Badi, Tarhini & Khan, 2018; Bennett, 2017). Those with authority alternatively are also accountable (Benfeldt, Persson & Madsen, 2019) for ensuring data governance is implemented and adhered to by others in the organisation.

Accountability for data governance includes conforming to regulations, standards and internal policies (Abraham, vom Brocke & Schneider, 2019; Al-Badi, Tarhini & Khan, 2018; Bennett, 2017, Chukwarah et al., 2024). Internal policies are particularly important for governing human behaviour within an organisation. Human behaviour demonstrating appropriate management and use of data is an expected outcome of positive and effective data governance (Benfeldt, Persson & Madsen, 2019). To determine if accountability for data

governance has been achieved, data governance management should be monitored for compliance to policies and regulations (Abraham, vom Brocke & Schneider, 2019).

2. Shedding a Light on Data

Data that organizations collect is either structured or unstructured in nature where the latter is referred to as dark data (Ajis & Baharin, 2019; Imdad et al., 2020). Structured data can be accessed and analyzed using data mining tools as it is stored in relational databases (Imdad et al., 2020). This data can be categorized as business-critical data being important to the organization or redundant, obsolete and trivial (ROT) data. ROT data has little or no value to an organization as it is either duplicated or once useful but is now no longer required (Imdad et al., 2020).

The alternative data type is unstructured data. It is estimated that unstructured data (dark data) represents an estimated 70% to 90% of all enterprise data (Mesaglio & LeHong, 2024). Unstructured data is increasingly complex and includes instant messaging, short message service text, electronic mail, spreadsheets, XML files, voice, video, and social media interactions (Baker and Sjoberg, 2018; Henderson, 2017) and does not have a pre-defined model or structure (The University of Queensland, 2021; Henderson, 2017). Unstructured data has the characteristics of unstructured information or meaning, making its analysis difficult as the content and data value is buried inside the unstructured data requiring human interpretation (Ajis, Zakaria & Ahmad, 2022; Ajis & Baharin, 2019). The inability to access data puts the accuracy of the current approach used for data analysis at risk, which may lead to poor decision making based on incomplete or incorrect data (Ajis, Zakaria & Ahmad, 2022).

An additional issue with dark data is any negative perceptions and beliefs of those in the company or organization. Dark data may be considered in some organizations as too old to be of value, or of limited value because it is in a format that cannot be accessed with available systems or tools (Edwards et al., cited in Azeroual, Nikiforova & Sha, 2023). Generally, it is not even known whether these data exist (Azeroual, Nikiforova & Sha, 2023). However, Gimpel (cited in Azeroual, Nikiforova & Sha, 2023) says dark data can represent one of an organization's greatest untapped resources, especially when data are increasingly becoming an important asset. Data governance provides the ability build trust in data providing guidance on how data is collected, analyzed, published or used (Eryurek et al., 2021; Ajis & Baharin, 2019).

3. Method

A comprehensive literature review approach was implemented using multiple databases including Science Direct database, Web of Science, Scopus, Proquest and Google Scholar. Key words found in the title and or abstract of published articles were included in the inquiry. Key words used to find the data were “dark data” OR “secondary use of data” OR “big data” OR “metadata” AND “governance” OR “data governance”. Commencing the search for suitable sources, included the criteria of using empirical studies. This criterion was extended to include other types of articles due to dearth of research. Discussion papers, expert opinions and all literature reviews were also included to ensure collection of data about concepts, procedures and strategies useful for contemporary data governance knowledge. At all stages of data collection and extraction, discussions, deliberations and consensus were achieved between both authors to ensure rigour to this inquiry (Paul and Criado, 2020).

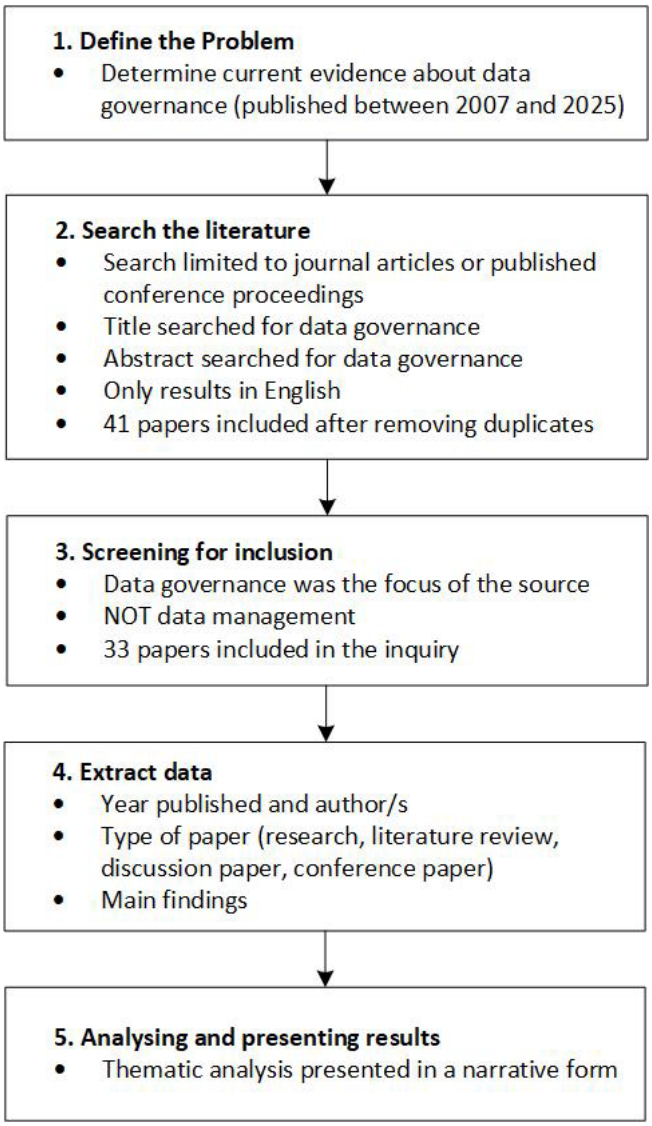

Figure 1: Flow chart of Research Method

4. Findings/Results

This inquiry implemented a comprehensive literature review with a final sample size of 33 published articles included in this study. This inquiry examined governance for secondary data (dark data). Key concepts identified that were important to data governance of dark data were data quality, data security, and data access.

5. Discussion

The viewpoint that data is a valuable resource has grown increasingly prevalent among business and Information Technology (IT) executives. Dark data is now seen as a business opportunity for further use of data, however there are challenges to work with dark data. Ashbel (2021) states that data governance, data compliance, and cost-efficient storage are three such challenges. Data governance is a set of policies and practices that an organization can establish to support their data management (Almeida, 2021) creating clear visibility into an organisation's stored data. Policies and procedures that govern data management will help eliminate inconsistencies from dark data (Azeroual, Nikiforova & Sha, 2023) and improve data quality, thus helping to reduce data management costs (Ashbel, 2021).

6. Data Quality

Data quality is key to effective decision making (Azeroual, Nikiforova & Sha, 2023; Black et al., 2023; Chukwurah et al., 2024; Gunawong, 2023). High-quality data can generate insights, promote knowledge and

contribute to business outcomes (Chukwurah et al., 2024). Alternatively, duplicated and other deficient data content, typically hidden in dark data can lead to poor data quality, creating inaccurate data leading to ill-informed decisions, loss of competitive strength and at worse, reputation damage (Azeroual, Nikiforova & Sha, 2023; Ajis & Baharin, 2019).

Contributing to data quality issues have been highlighted in the literature such as massive volumes and variety of data, high velocity, and various automatically generated data (Azeroual, Nikiforova & Sha, 2023). Poor data can be inconsistent, inaccurate, incorrect, or incomplete (Azeroual, Nikiforova & Sha, 2023; Gunawong, 2023) which usually also leads to the existence of duplicates, which in turn constitute data contamination (Azeroual, Nikiforova & Sha, 2023). Logically, dark data and duplicates, given their nature, need to be considered from different perspectives. Not only technological solutions are needed that would reveal poor- or low-quality data, but also organizational and cultural changes related to data management and governance within the organization will be necessary for increasing future use of dark data (Azeroual, Nikiforova & Sha, 2023).

As the volume of data continues to grow, further compounded by multiple sources of data, this will lead to increased storage costs to an organisation (Ajis & Baharin, 2019). This situation also increases duplication and along with non-unique data degrades data quality, resulting in inefficient and inaccurate data-based decisions (Azeroual, Nikiforova & Sha, 2023).

To mitigate the risks of poor data quality and to implement a proactive solution, data governance should commence with developing a data quality strategy (e.g., Abraham, vom Brocke & Schneider, 2019; European Foundation for Quality Management, 2011), including identifying roles and responsibilities definitions, and the reason for data quality management processes (e.g., Abraham, vom Brocke & Schneider, 2019; Malik 2013). The management of data quality should extend to data quality assessment (Azeroual, Nikiforova & Sha, 2023), the definition of data quality metrics (e.g., Abraham, vom Brocke & Schneider, 2019; Brous, Janssen & Vilminko-Heikkinen, 2016) and consistent measurement of data quality levels (e.g., Abraham, vom Brocke & Schneider, 2019; Azeroual, Nikiforova & Sha, 2023; Henderson 2017). The data quality strategy should consider how to achieve accurate, timely, complete and credible data (Gunawong, 2023). This governance strategy should also include how to manage and promptly address any data quality issues (Abraham, vom Brocke & Schneider, 2019; Henderson 2017) in a timely manner (Brous & Janssen, 2020).

Data governance establishes data management processes which can manage data quality (Azeroual, Nikiforova & Sha, 2023; Brous & Janssen, 2020). Azeroual, Nikiforova & Sha, (2023) suggesting data governance developing policies with transparent procedures to ensure quality assurance and management processes (e.g., data cleansing) and employment of Artificial Intelligence. Such procedures are designed to improve understanding of the data to use them more effectively and efficiently. Data governance is therefore able to address the challenge of data quality that is inconsistent, inaccurate, incorrect, or incomplete reversing this situation to achieve data accuracy, availability, completeness, consistency, and timeliness of data (Abraham, vom Brocke & Schneider, 2019; Azeroual, Nikiforova & Sha, 2023).

7. Data Security and Data Access

Data security is the practice of safeguarding organizational information against theft, cyberattacks, and data breaches due to unauthorized access (Wong & Norbaini, 2024). Dark data typically has widespread duplicated content without explicit oversight which weakens security (Ajis & Baharin, 2019). Data sharing is increasing both internally and externally to enhance transparency, inclusivity and productivity (Chukwurah et al., 2024). With the use of big data, hackers have multiple potential entry points leading to leaked data, lost, stolen, or breached dark data. To stop intruders from accessing or stealing data, security measures must be in place at all sites where sensitive data is stored (Wong & Norbaini, 2024). If security is not controlled damaged reputations as well as loss of competitive advantage for the organization can result (Ajis & Baharin, 2019). Data governance to enhance security is essential for protecting the company's data.

Protecting the company's data is an evolving process requiring the organisation to be agile as changes in IT and data requirements continue to occur. Contemporary business strategies to enhance productivity include examples like streamlining IT systems for greater transparency, end-user ease and sharing data. Sharing data occurs both internally between company departments as well as with external users such as third-party organisations. With the modern trend of increasing expectations for data management to remain agile, securing systems and in particular protecting data has become more complicated and requiring sound data governance.

Security measures must be incorporated into a data governance strategy to guarantee that no one has more access than is necessary to perform their duties, even at the data field level (Khatri and Brown, 2010; Wong & Norbaini, 2024). The early work of Khatri and Brown (2010) continues to be relevant when they argued that a data governance framework was essential to address data security, which included data access. Data governance should commence with risk analysis to identify data needs for different business departments ensuring the 'need for access' (e.g. Abraham, vom Brocke & Schneider, 2019; Khatri and Brown, 2010) to protect confidentiality and privacy. The need for access includes identifying who, when, how, and what access is permitted as well as who will be responsible for ensuring these guidelines are met (Azeroual, Nikiforova & Sha, 2023). Such data governance establishes the right of access and accountability for the process for securing the data, starting when data is created, and transferred between digital systems to where and to whom has the authority to access this data (both internally and externally) (Wong & Norbaini, 2024).

Data governance should align with laws and regulations. Appropriate security controls should abide by laws and regulations like the General Data Protection Regulation (GDPR) and the California Privacy Rights Act (CPRA) (Wong & Norbaini, 2024). By adhering to laws and regulations companies avoid non-compliance, enhance data quality, and reduce the risk of data being breached (Chukwurah et al., 2024). Regulations provide the foundation for policies, as compliance is essential. Complying to laws and regulations provides guiding principles that should be incorporated when writing and updating organisation's policies and procedures as well as monitoring for auditing purposes. This includes examining the use of policies and reported data issues and response to such issues (Abraham, vom Brocke & Schneider, 2019; Brous & Janssen, 2020; Palczewska et al. 2013).

A dependable and uniform structure provided by clearly articulated policies and procedures for handling and using data is the desired outcome from data governance policies (e.g. Abraham, vom Brocke & Schneider, 2019; Khatri and Brown, 2010; Wong & Norbaini, 2024). Policies for data governance aims to guarantee data security, accuracy, access and consistency of handling the data (Brous & Janssen, 2020; Wong & Norbaini, 2024).

Those authoring policies should aim for transparency of expectations for accessing and handling data appropriately both within the organisation and with external stakeholders (Chukwurah et al., 2024; Wong & Norbaini, 2024). Data governance through policies should apply to all facet of data management across the data lifecycle. This includes production including data quality; storage; upkeep; usage including access and appropriate use; and disposal of data (e.g. Eryurek et al., 2021; Wong & Norbaini, 2024).

Further detail required in data governance policies when establishing data security controls in data governance (Abraham, vom Brocke & Schneider, 2019; Tallon, Ramirez & Short, 2014) should include data security roles (e.g. Abraham, vom Brocke & Schneider, 2019; Khatri and Brown, 2010) and applying industry and government standards (e.g. Abraham, vom Brocke & Schneider, 2019; Khatri and Brown, 2010). Finally, the data governance policy for data access should outline the organisation's response plan if data is accessed without authorisation or used inappropriately (Wong & Norbaini, 2024).

Timely addressing data breaches starts with clear procedures of what one should do to minimise possible hazards or an actual breach (Chukwurah et al., 2024; Wong & Norbaini, 2024). Certainly, companies that use data governance to implement risk management, put security measures in place and continue to monitor and identify any potential or actual breach will ensure that sensitive data is protected by lowering the risk of data breaches (Chukwurah et al., 2024; Wong & Norbaini, 2024). A strong data governance framework should specify what should be done to minimize harm and stop similar incidents in the future. Reports should be expected as part of the governance approach showing accountability (Chukwurah et al, 2024) and action to avoid further data management issues.

8. The Key to Successful Data Governance

Successful implementation of data governance frameworks depends on interrelated sets of factors. The underpinning factors commence with developing transparent and robust data governance policies (Ajis & Baharin, 2019) and customising the governance framework for things like big data (Kim and Cho, 2018). Addressing specific technological demands such as tailored quality standards and big data algorithms (Chandra et al. (2023) means that governance provides a purposeful and useful framework for managing the data. Along with solid quality data improvement, integrity and security controls are essential (Henriques et al. 2021). Security controls should include ensuring competencies in data privacy (Henriques et al. 2021) such as avoiding leaks of personal information (Kim & Cho, 2018). Other integrated factors that complement foundational

factors are data governance aligning with business goals, organizational structure, human and culture. Technical implementation factors all influence success of data governance.

Sound data governance considers how people impact on implementation of the governance framework and the resultant data management. The first organisational structure, human and culture factor to consider is employee perceptions of the value of using dark data and the trust in the resultant data driven decision (Alhassan, Sammon & Daly, 2019; Ajis & Baharin, 2019). Organisational factors need to provide guidelines of when and who should be clearly identified as the owner and user of the data in the governance framework (Brous & Janssen, 2020). When considering such detail, one should not forget to engage stakeholders in the process of developing a governance framework. Stakeholders are likely to be people both internal and external to the company. Data governance is likely to impact on both parties.

9. Conclusion

Increased productivity and positive business outcomes can be gained by organisations searching out dark data and optimising data use. Regardless of the benefits of using dark data, inaction to use data may be detrimental to a company working with incomplete datasets. As other companies become more efficient with data use, productivity is gained through the use of unused data resulting in greater accuracy of data. Those companies who are embracing dark data are likely to have a competitive edge. Others yet to embrace dark data will need solid data governance to assist with the conversion of dark data to useful data and gain quality datasets. When handling and using dark data, the original providers of the data (consumers, participants) should be protected, and data secured to protect the company. Data is used by people and as such organisational structure and culture factors are key to robust data governance. Dark data can potentially be a hidden gem and should be unlocked; but cannot be mined appropriately without data governance.

Ethics declaration: I did not need an ethical clearance for the research referred to in this paper.

AI declaration: I did not use an AI tool in the development of this paper.

References

Abraham, R., vom Brocke, J., & Schneider, J. (2019). Data governance: A conceptual framework, structured review, and research agenda.

Ajis, A. F. M., Zakaria, S., & Ahmad, A. R. (2022). Demystifying Dark Data Characteristics in Small and Medium Enterprises: A Malaysian Experience.

Ajis, A. F. M., & Baharin, S. H. (2019). S.H. Dark Data Management as frontier of Information Governance.

Al-Badi, A., Tarhini, A., & Khan, A. I. (2018). Exploring Big Data Governance Frameworks.

Alhassan, I., Sammon, D., & Daly, M. (2016). Data governance activities: an analysis of the literature.

Almeida, C. A., et al. (2021). Excavating FAIR Data: the Case of the Multicenter Animal Spinal Cord Injury Study (MASCIS), Blood Pressure, and Neuro-Recovery.

Ashbel, A. (2021). Dark Data A Challenge Enterprise Data Management Can't Ignore.

Azeroual, O., Nikiforova, A., & Sha, K. (2023). Overlooked Aspects of Data Governance: Workflow Framework For Enterprise Data Deduplication.

Baker, S., & Sjoberg, P. (2018). Intelligent Data Governance, John Wiley & Sons Inc., Hoboken, New Jersey, USA.

Benfeldt, O., Persson, J. S., & Madsen, S. (2019). Data Governance as a Collective Action Problem.

Bennett, S. (2017). What is information governance and how does it differ from data governance?

Black, S. et al. (2023). Data governance and the secondary use of data: The board influence.

Brous, P., Janssen, M., & Vilminko-Heikkinen, R. (2016). Coordinating Decision-Making in Data Management Activities: A Systematic Review of Data Governance Principles.

Brous, P, & Janssen, M. (2020). Trusted Decision-Making: Data Governance for Creating Trust in Data Science Decision Outcomes.

Chandra, Y. U. et al. (2023). Control and Data Integrity are Important Factors of Data Governance Technology, 2023 10th International Conference on ICT for Smart Society (ICISS), 06-07 September 2023, DOI: 10.1109/ICISS59129.2023.10291563

Cheong, L. K., & Chang, V. (2007). The Need for Data Governance: A Case Study.

Chukwurah, N. et al. (2024). Strategies for engaging stakeholders in data governance: Building effective communication and collaboration.

Duarte, F. (2024). Amount of Data Created Daily (2024). Available at: https://explodingtopics.com/blog/data-generated-per-day (Accessed: 30 January 2025).

European Foundation for Quality Management. (2011). Framework for Corporate Data Quality Management.

Eryurek, E. et al. (2021). Data Governance: The Definitive Guide. People, Processes, and Tools to operationalize Data Trustworthiness, O'Reilly Media Inc., Sebastopol, California, USA.

Gunawong, P. (2023). Data governance for wicked problems: A case from the Thai health system.

Henderson, S. (2017). DAMA-DMBOK: Data Management Body of Knowledge, 2nd. Edition. Technics Publications, Basking Ridge, New Jersey, USA.

Henriques et al. (2021). An automated closed-loop framework to enforce security policies from anomaly detection, Computers & Security, Volume 123, December 2022, 102949, Available at: https://doi.org/10.1016/j.cose.2022.102949 (Accessed on 30 January 2025).

Imdad, M. et al. (2020). Dark Data: Opportunities and Challenges.

Khatri, V., & Brown, C. V. (2010). Designing data Governance, Communications of the ACM, vol. 53 | no. 1, DOI: 10.1145/1629175.1629210.

Kim, H. T., & Cho, J. S. (2018). Data governance framework for big data implementation with NPS Case Analysis in Korea, Journal of Business and Retail Management Research (JBRMR), Vol. 12 Issue 3, Available at: https://doi.org/10.24052/JBRMR/V12IS03/ART-04 (Accessed on 30 January 2025)

Malik, P. (2013). Governing Big Data: Principles and practices. IBM Journal of Research and Development, Volume 57, Issue 3/4, pp. 1-13.

Mesaglio, M., & LeHong, H. (2024). CIOs: Your AI Tech Stack Needs a New Look, Available at: https://www.gartner.com/en/articles/ai-tech-stack (Accessed on 30 January 2025).

Palczewska, A. et al. (2013). Towards a model governance in predicting toxicology, International Journal of Information Management. 33(3), 567-582.

Paul, J., & Criado, A. R. (2020). The art of writing literature review: What do we know and what do we need to know? International business review, 29, 1010717, Available at: https://doi.org/10.1016/j.ibusrev.2020.101717 (Accessed on 30 January 2025).

Tallon, P. P., Ramirez, R. V., & Short, J. E. (2014). The information artifact in IT governance: toward a theory of information governance.

The University of Queensland. (2021). Data Governance Essentials Handbook, Available at: https//data.uq.edu.au/files/6833/Data (Accessed on 30 January 2025).

Wong, H. M., & Norbaini, S. F. (2024). The Data Governance: A Comprehensive Literature Review from Professional Viewpoints.