A HOLISTIC PERSPECTIVE ON DATA GOVERNANCE

Anacleto Correia *, Pedro B. Água *

* CINAV, Naval School, Military University Institute, Almada, Portugal

How to cite: Correia, A., & Água, P. B. (2021). A holistic perspective on data governance. In S. Hundal, A. Kostyuk, & D. Govorun (Eds.), *Corporate governance: A search for emerging trends in the pandemic times* (pp. 69–75). https://doi.org/10.22495/cgsetpt12

Received: 26.04.2021
Accepted: 30.04.2021
Keywords: Data Governance, Corporate Governance, Data Quality, Data Management, Enterprise Architecture
JEL Classification: M15
DOI: 10.22495/cgsetpt12

Abstract

Data governance sets the principles and rules organizations should follow for the effective use of data. Organizations also expect by means of adequate data governance the attainment of cost-effective and lower-risk operations. Despite data governance awareness in recent years, there is a lack of a holistic view of the organization's data governance that could help both practitioners and researchers to have an overall map of the current situation and anticipate the further steps needed to raise its level of maturity. This exploratory research proposes a classification scheme for data architecture according to two orthogonal dimensions: the perspective of stakeholders (from corporate board to end-users) as well as the primitives that contribute to better data governance. The proposed scheme, evolved from enterprise architecture research, is in line with other solutions aimed at aligning the business and IT within organisations.

1. INTRODUCTION

The amount of data produced every day by organizations is overwhelming. More than 59 zettabytes of data were expected to be handled in 2020. The rate of data increase is so that 90% of all data was created in the last two years (IDC, 2020). Although, until now, the most relevant data stored in organizations are stored in databases, running e-commerce, enterprise resource planning (ERP) systems, and email, it is expected that unstructured data will become prevalent, including besides traditional office documents, video and audio files, as well as geospatial data, Internet of things (IoT) data, and streaming. Each year an increasing amount of entertainment video is produced and consumed. Hundreds of billions of IoT sensors are being embedded all over the places, generating increasing amounts of data along with metadata. The data processing speed and bandwidth are accelerating data transfer and reducing latency. Supported by satellites, 5G and 6G networks, the finding of new tools for creating, sharing, and consuming data, and the steady addition of new data producers and consumers ensure the hungry for increasing data will growth persistently (Press, 2020). This trend is escalating the use of cloud storage and computation to a point that, by 2025, it is expected that around 50% of all data will be stored in the cloud (IDC, 2020).

The more the amount of data the more the concerns regarding data governance to ensure its integrity and accessibility. Some of the most recognised financial scandals (e.g., Enron, WorldCom, Lehman Brothers), were based on data perversion to hide billions of dollars of bad debt and loans, inflation of earnings or assets through accounting loopholes. On the other hand, data breaches affecting well-known organizations (e.g., eBay, LinkedIn, Yahoo, Facebook) are becoming far too common. Privacy of billions of users has been compromised since personal data stolen from breaches (e.g., credit card numbers, email addresses, personal photos, passwords) were made publicly available or put up for sale on the dark web (Swinhoe, 2021). In both cases, corporate boards are being held responsible either for the accuracy of the organisation's financial data (Cheong & Chang, 2007) and for data leakage compromising stakeholders' privacy.

An understanding of the data governance role is crucial for corporate governance to nurture data quality and protection, as well as other ones such as data fusion from several sources and the integration of several kinds of systems and applications (e.g., IoT devices, ERP, Data Analytics, Big Data) (Cheong & Chang, 2007). One of the several definitions of data governance is "the exercise of authority, control, and shared decision making over the management of data assets" (Brous, Janssen, & Vilminko-Heikkinen, 2016). Through adequate data governance, organizations are equipped to ensure that data are managed appropriately, providing for people at different levels of decision with the right information needed at the right moment (Thompson, Ravindran, & Nicosia, 2015).

This study proposes a classification scheme, which provides a holistic view on data architecture and allowing that the right actions can be triggered to correct non-conformities on data management or even raise the level of maturity of data governance.

2. BACKGROUND

In the last decades of the previous century, many organizations recognized the relevance for creating the data administration (DA) function under the supervision of the corporate resources of information (Holloway, 1986). The relevance of this function anticipated the nowadays importance given to data governance. The role of data administration was to promote the planning and coordination of the information resource usage across organization, among related applications and business areas. By doing so, data sharing could be maximized and data redundancy minimized. Data administrators make data sharable and consistent across applications by using logical data modelling. They ensured that several other tasks were performed, as for instance: gathering business requirements, requirements analysis, business modelling based on requirements, definition and enforcement of standards and conventions regarding names and terms, collecting users' data definition, management and stewardship of the metadata repository and data modelling tools. Furthermore, DA supported the technical function of database administration on creating physical databases from logical models.

Another perspective highlighting the relevance of data was given by enterprise architecture (EA) frameworks (Pieterse, 2015). The relevance of data as one of the building blocks of enterprise architecture was highlighted by Zachman (1999) and Sowa and Zachman (1992). Currently, the TOGAF (TOGAF, 2018), one of the most widely used EA frameworks, describes a detailed method for developing, within enterprise architecture, data architecture as one of its parts.

Weill and Ross (2004) define data governance as a framework for decision rights and accountabilities to encourage desirable behaviour in the use of data (Ross & Weill, 2004). On the other hand, the Data Management Association (http://www.dama.org/) provides a practitioner's perspective and, besides considering the relevance of the specification of a framework, also highlights the practices surrounding the data governance process. Nevertheless, it seems consensual that the important goals of data governance are: 1) provide conditions for better decision making, 2) support regulatory compliance and risk reduction regarding data privacy & security, 3) raise business performance, 4) support business integration, and 5) increase IT-business alignment (Thompson et al., 2015).

According to Brous et al. (2016), little evidence has been found indicating what actually has to be organized under data governance and what data governance processes may entail. On the other hand, most research has focused on structuring or organizing data governance, the data governance processes to be implemented and data governance coordination. The suggested proposal in the next section intends to systematize the information found in the literature about data governance.

3. PROPOSAL

In this work, using the concept of symmetry, the classification scheme proposed by Zachman (1999) for enterprise architecture, is applied to data governance. The rationale for this analogy is grounded in the realization that data governance (as part of data architecture) is also part of the enterprise architecture. Therefore, for sake of symmetry, it is required the parity of relevant characteristics of these parts to compose the whole of enterprise architecture.

The concept of symmetry in architecture is ancient. According to Roman architect Vitruvius, symmetry consists of the union and conformity of the parts of a work, in relation to its totality. Symmetry also derives from the Greek concept of analogy, which is understood as the relationship between all parts of a structure with the whole structure. That is why a uniform symmetry between data architecture (and data governance) and enterprise architecture is required. In general, uniform symmetry occurs in architecture when the same motif reigns throughout the structure.

The proposed classification scheme for data governance (Table 1), based on the Zachman's framework, is depicted as a two-dimensional matrix composed by: 1) rows as top-down *perspectives* of data, from contextual corporate board perspective to end-users' operations perspective, and 2) columns as *primitive* concepts, triggered by interrogative adverbs. Each perspective in the first dimension aims at a *target* (i.e., the reification of abstract ideas into instantiation), labelled as *Identification*, *Requirements*, *Representation*, *Specification*, *Configuration*, and *Instantiation*. Each one of the reification levels corresponds to a different organizational level with different perspectives of their role in what concerns data: *Governance*, *Management*, *Modelling*, *Building*, *Implementing*, and *Using*. The second dimension intends at the elicitation of a certain type of artifacts built in response to specific adverbs: *Inventory* (What), *Process* (How), *Distribution* (Where), *Responsibility* (Who), *Timing* (When), and *Motivation* (Why). Each column elicits artifacts derived from the following primitive concept: *Sets*, *Flows*, *Networks*, *Assignments*, *Cycles*, and *Intentions*. The final classifications are depicted in the cells resulting from the intersection between the perspectives and the concepts, and representing the tools used for data governance. The overall matrix constitutes the total set of descriptive representations that are relevant for describing any architectural part of an organization, in particular the data architecture, as well as the overall organization itself. The classification scheme as a classification structure is presented in Table 1.

Table 1. Classification scheme for data governance

Classification	*Inventory (What)*	*Process (How)*	*Distribution (Where)*	*Responsibility (Who)*	*Timing (When)*	*Motivation (Why)*	*Target*
Perspective							
Governance	Data sufficiency	Regulatory compliance	Centralized	Corporate board	Optimized	Information management strategy	Identification
Management	Data principles and rules	Cost & productivity	Consultative	Chief data officer	Managed	Data quality	Requirements
Modelling	Meta data	Change management	Balanced	Stewardship	Defined	Metadata strategy	Representation
Building	Data repositories	Fusion & integration	Federated	Repositories' supervisors	Repeatable	Data security & privacy	Specification
Implementing	Data integrity	Extract, transform & load	Independent	IT technicians	Initial	Data access performance	Configuration
Using	Unstructured & structured data	Decision support & CRUD	Transparency	End-users	UpToDate	Data effectiveness	Instantiation
Primitive	Sets	Flows	Networks	Assignments	Cycles	Intentions	

One can detail, in terms of data governance, the *perspective* dimension by describing its different levels of abstraction, specifically: 1) *Governance* addresses the role of corporate board directors regarding the organizations’ strategy to the data asset; 2) *Management* concerns with the definition of principles and rules for data management; 3) *Modelling* provides guidelines for standardized use of data; 4) *Building* relates with the technical creation and maintenance of data repositories; 5) *Implementing*, also a technical perspective, concerns on how to make data available to the end-users; and 6) *Using* as a perspective that represents how the organization makes use of data to accomplish the strategy and objectives.

A more detailed explanation of the matrix requires the description of the meaning of each cell under the classification scheme. Due to limitation of space, only the cells of the *Inventory* column are described, which focuses on how data assets can be approached by the decreasing level of abstraction of perspectives:

- *data sufficiency* — corporate boards should identify the organization’s data perimeter, i.e., the extension at which the organization should capture and store data, avoiding handling unnecessary data, or incurring in situations of data privacy abuse having as a consequence the possible data leakage that could harm, financially and reputationally, the organization;
- *data principles and rules* —data management should define the guidelines and constraints regarding the storage of data by the organization;
- *meta data* — business analysts contribute with the elicitation of the data attributes. The metadata information can be of several types

including the description of assets (descriptive metadata); description of data containers characterizing how compound objects are put together (structural metadata); information about resources' management (administrative metadata); contents and quality of statistical data (reference metadata); description of the processes for collecting, processing, or producing statistical data (statistical metadata); and information about the legal owner (legal metadata). Standards (e.g., ISO/IEC 11179-1:2015) and tools can be used for standardization of the metadata;

- *data repositories* — IT supervisors, based on the metadata specifications, define and maintain the organization's physical repositories of data (e.g., controlled vocabularies, taxonomies, thesauri, data dictionaries, metadata registries);
- *data integrity* — developers configure business rules and data constraints in applications and databases. Reference data should be used to validate data entries by defining the set of permissible values for data fields, preferably based on values defined by standards organizations;
- *unstructured and structured data* — data required for conducting business operations and support decision-making at different organizational levels. Structured data reside within pre-defined models (e.g., relational databases, master data files, spreadsheets), while unstructured data is not supported by pre-defined data models (e.g., e-mails, pictures, audio, video, scanned documents).

4. CONCLUSION

Organizations produce and use an immense amount of data. As an organization, strategic asset data must be appropriately governed at all institutional levels (perspective) and characterized in accordance with uncovered relations with other (primitive) concepts. In this research, we propose a classification scheme for data governance. The tool, derived from an enterprise architecture framework, intends to be a way to align data governance strategy with the enterprise architecture. As future work, we intend to develop the proposed model by deepening the relationships between data governance and enterprise architecture.

REFERENCES

1. Brous, P., Janssen, M., & Vilminko-Heikkinen, R. (2016). Coordinating decision-making in data management activities: A systematic review of data governance principles. In H. J. Scholl, O. Glassey, M. Janssen, B. Klievink, I. Lindgren, P. Parycek,… D. Sá Soares (Eds.), *Electronic Government: 15th IFIP WG 8.5 International Conference, EGOV 2016*. (pp. 115–125). https://doi.org/10.1007/978-3-319-44421-5_9

2. Cheong, L. K., & Chang, V. (2007). The need for data governance: A case study. In *ACIS 2007 Proceedings — 18th Australasian Conference on Information Systems*. Retrieved from https://core.ac.uk/download/pdf/301346974.pdf
3. Holloway, S. (1986). Data administration in the organization: What is it, who does it and why? *Data Processing, 28*(4), 195–198. https://doi.org/10.1016/0011-684X(86)90361-8
4. IDC. (2020, May 8). *IDC's Global DataSphere forecast shows continued steady growth in the creation and consumption of data*. Retrieved from https://bit.ly/3wRXbpp
5. Pieterse, R. (2015). *Enterprise architecture frameworks, methods and tools*. Morrisville, NC: Lulu Press.
6. Press, G. (2020, January 6). 6 predictions about data in 2020 and the coming decade. *Forbes*. Retrieved from https://bit.ly/3sn7nD3
7. Sowa, J. F., & Zachman, J. A. (1992). Extending and formalizing the framework for information systems architecture. *IBM Systems Journal, 31*(3), 590–616. Retrieved from http://www.jfsowa.com/pubs/sowazach.pdf
8. Swinhoe, D. (2021, January 8). The 15 biggest data breaches of the 21st century. *CSO*. Retrieved from https://bit.ly/3g7D68M
9. Thompson, N., Ravindran, R., & Nicosia, S. (2015). Government data does not mean data governance: Lessons learned from a public sector application audit. *Government Information Quarterly, 32*(3), 316–322. https://doi.org/10.1016/j.giq.2015.05.001
10. TOGAF. (2018). *Welcome to the TOGAF® Standard, Version 9.2, a standard of The Open Group*. Retrieved from the Open Group website: https://pubs.opengroup.org/architecture/togaf9-doc/arch/
11. Weill, P., & Ross, J. W. (2004). *IT governance: How top performers manage IT decisions rights for superior results*. Brighton, MA: Harvard Business Review Press.
12. Zachman, J. A. (1999). A framework for information systems architecture. *IBM Systems Journal, 38*(2.3), 454–470. https://doi.org/10.1147/sj.382.0454