DOI: 10.1145/1629175.1629210

BY VIJAY KHATRI AND CAROL V. BROWN

Designing Data Governance

ORGANIZATIONS ARE BECOMING INCREASINGLY SERIOUS about the notion of "data as an asset" as they face increasing pressure for reporting a "single version of the truth." In a 2006 survey of 359 North American organizations that had deployed business intelligence and analytic systems, a program for the governance of data was reported to be one of the five success "practices" for deriving business value from data assets.[a] In light of the opportunities to leverage data assets as well ensure legislative compliance to mandates such as the Sarbanes-Oxley (SOX) Act and Basel II, data governance has also recently been given significant prominence in practitioners' conferences, such as TDWI (The Data Warehousing Institute) World Conference and DAMA (Data Management Association) International Symposium.

The objective of this article is to provide an overall framework for data governance that can be used by researchers to focus on important data governance issues, and by practitioners to develop an effective data governance approach, strategy and design. Designing data governance requires stepping back from day-to-day decision making and focusing on identifying the fundamental decisions that need to be made and who should be making them. Based on Weill and Ross,[10] we also differentiate between governance and management as follows:

- *Governance* refers to what decisions must be made to ensure effective management and use of IT (*decision domains*) and who makes the decisions (*locus of accountability for decision-making*).
- *Management* involves making and implementing decisions.

For example, governance includes establishing who in the organization holds decision rights for determining standards for data quality. Management involves determining the actual metrics employed for data quality. Here, we focus on the former.

Corporate governance has been defined as a set of relationships between a company's management, its board, its shareholders and other stakeholders that provide a structure for determining organizational objectives and monitoring performance, thereby ensuring that corporate objectives are attained. Considering the synergy between macroeconomic and structural policies, corporate governance is a key element in not only improving economic efficiency and growth, but also enhancing corporate confidence.[b] A framework for linking corporate and IT governance (see Figure 1) has been proposed by Weill and Ross.[10]

Unlike these authors, however, we differentiate between IT assets and information assets: *IT assets* refers to technologies (computers, communication and databases) that help support the automation of well-defined tasks, while *information assets* (or data) are defined as facts having value or potential value that are documented. Note that in the context of this article, we do not differentiate between data and information.

Next, we use the Weill and Ross framework for IT governance as a starting point for our own framework for data governance. We then propose a set

a http://mediakit.businessweek.com/pdf/research/KnightsbridgeWhitePaper.pdf (last viewed on August 2, 2007)

b http://www.oecd.org/dataoecd/32/18/31557724.pdf

of five data decision domains, why they are important, and guidelines for what governance is needed for each decision domain. By operationalizing the *locus of accountability of decision making* (the "who") for each decision domain, we create a data governance matrix, which can be used by practitioners to design their data governance. The insights presented here have been informed by field research, and address an area that is of growing interest to the information systems (IS) research and practice community.

IT Governance as the Context for Data Governance

IT governance refers to who holds the decision rights and is held accountable for an organization's decision-making about *IT assets*. In their IT governance framework, Weill and Ross propose that governance design includes five major decision domains: IT principles; IT architecture; IT infrastructure; Business application needs; and IT investment and prioritization. Although the five key decisions are interrelated, each of these decisions deals with a distinctive set of core issues. *IT principles* clarify the role that IT plays in the organization and drive the IT architecture decisions that establish the *IT infrastructure*. The organization's IT infrastructure capabilities enable its *business application needs*, and the need for new IT applications can create new IT infrastructure requirements. *IT investment and prioritization* decisions are in turn shaped by the organization's IT principles, architecture, infrastructure, and application needs.

Data Governance: The Five Decision Domains

Data governance refers to who holds the decision rights and is held accountable for an organization's decision-making about its *data assets*. Our framework for data governance includes five interrelated decision domains: Data principles; Data quality; Metadata; Data access; and Data lifecycle. Figure 2 emphasizes the interconnections between these decision domains. *Data principles*, shown at the top of the framework, establish the direction for all other decisions. An organization's data principles set the boundary requirements for the intended uses of data, which set the organization's standards for *data quality*, which in turn are the basis for how data is interpreted (*metadata*) and accessed (*data access*) by users. Decisions that define the production, retention and retirement of data (*data lifecycle*) play a key role in operationalizing the data principles into IT infrastructure.

Table 1 summarizes the scope of

Figure 1: Key organizational assets to be governed; adapted from Weill and Ross.[10]

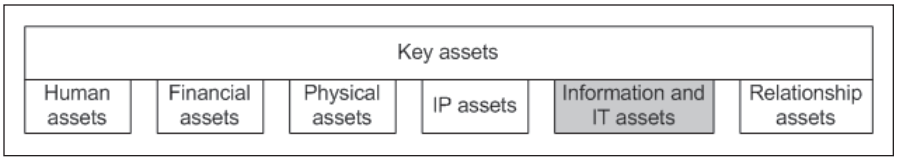


Figure 2: Decision domains for data governance.

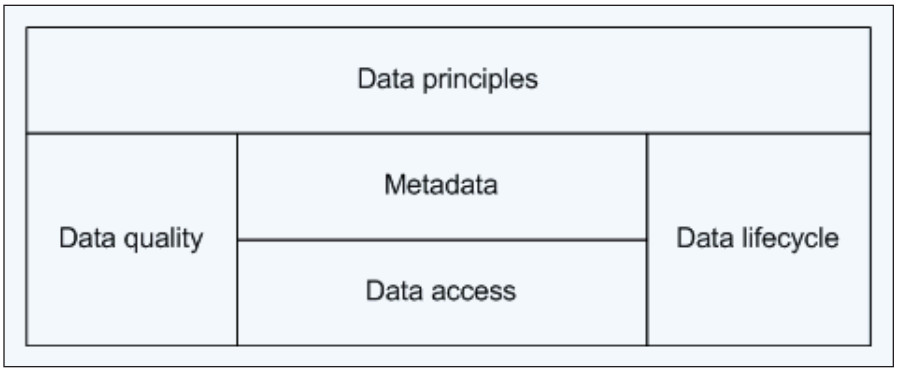
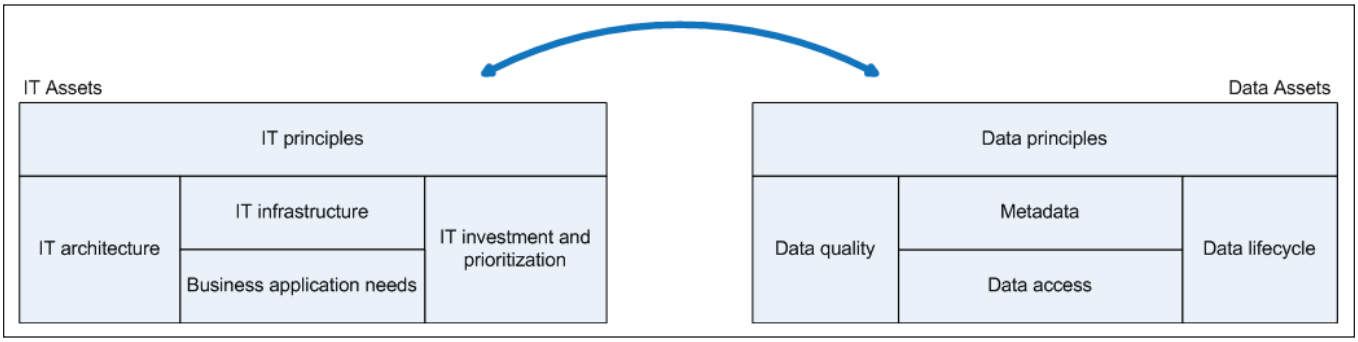


Table 1: Framework for data decision domains.

Data Governance Domains	Domain Decisions	Potential Roles or Locus of Accountability
Data Principles • Clarifying the role of data as an asset	• What are the uses of data for the business? • What are the mechanisms for communicating business uses of data on an ongoing basis? • What are the desirable behaviors for employing data as assets? • How are opportunities for sharing and reuse of data identified? • How does the regulatory environment influence the business uses of data?	• Data owner/trustee • Data custodian • Data steward • Data producer/supplier • Data consumer • Enterprise Data Committee/ Council
Data Quality • Establishing the requirements of intended use of data	• What are the standards for data quality with respect to accuracy, timeliness, completeness and credibility? • What is the program for establishing and communicating data quality? • How will data quality as well as the associated program be evaluated?	• Data owner • Subject matter expert • Data quality manager • Data quality analyst
Metadata • Establishing the semantics or "content" of data so that it is interpretable by the users	• What is the program for documenting the semantics of data? • How will data be consistently defined and modeled so that it is interpretable? • What is the plan to keep different types of metadata up-to-date?	• Enterprise data architect • Enterprise data modeler • Data modeling engineer • Data architect • Enterprise Architecture Committee
Data Access • Specifying access requirements of data	• What is the business value of data? • How will risk assessment be conducted on an ongoing basis? • How will assessment results be integrated with the overall compliance monitoring efforts? • What are data access standards and procedures? • What is the program for periodic monitoring and audit for compliance? • How is security awareness and education disseminated? • What is the program for backup and recovery?	• Data owner • Data beneficiary • Chief information security officer • Data security officer • Technical security analyst • Enterprise Architecture Development Committee
Data Lifecycle • Determining the definition, production, retention and retirement of data	• How is data inventoried? • What is the program for data definition, production, retention, and retirement for different types of data? • How do the compliance issues related to legislation affect data retention and archiving?	• Enterprise data architect • Information chain manager

Figure 3: Framework for IT and data decision domains.



each decision domain with examples of the types of decisions to be made for each domain. The far righthand column in Table 1 also provides examples of potential organizational roles that could be vested with decision rights for the various domains—that is, the "locus of accountability." A case study that we conducted with a large insurance company revealed several such roles for data governance: for example, the governance of data access was vested in an Enterprise Architecture Development Committee.

Data Principles. Effective data principles establish the linkage with the business. For example, the organizational decision to standardize business processes implies that there should be a clearly defined business owner of data assets (data principle). By delineating the business uses of data, data principles therefore establish the extent to which data is an enterprisewide asset, and thus what specific policies, standards and guidelines are appropriate. In keeping with the notion of data as an asset, data principles also establish/foster opportunities for sharing and reusing data. Each principle is supported by a rationale and a set of implications. Data principles take into account the usage of external data, such as, customer data from third-party service providers. An organization's data principles also take into consideration the regulatory environment that could influence the business uses of data.

Data principles therefore define the desirable behaviors both for IS professionals and business users. For example, the notion of business owners of data implies that business users have an important role in managing data quality as well as its lifecycle, interpretability and access. On the other hand, IS professionals play the role of data stewards wherein they employ IT tools (such as, DataFlux, Informatica Data Quality) that help surface quality issues for the business owners (or data owners/trustees).

Data Quality. Poor data quality can impact an enterprise at both operational and strategic levels[7]; current problems in data quality reportedly cost US businesses more than $611 billion every year in postage, printing, and staff overhead.[c] Similar to product quality,[3] the quality of data refers to its ability to satisfy its usage requirements.[5] While data quality has multiple dimensions, such as accuracy, timeliness, completeness and credibility, these dimensions are relative and need to be defined in the context of the end use of data.[1,5,9] For example, while 85% accuracy of the name, address, and phone number of physicians may be acceptable for an insurance company that is targeting physicians as potential customers, this metric would not be acceptable for organizations that need to notify prescribing physicians about a drug recall.

- *Accuracy* refers to correctness of data, that is, whether the recorded value is in conformity with actual value, with respect to its intended use.
- *Timeliness* indicates that the recorded value is up-to-date for the task at hand.
- *Completeness* suggests that the requisite values are recorded (not missing) and that it is of adequate depth/breadth.
- *Credibility* indicates the trustworthiness of the source as well as its content.

The data quality decision domain—which could be vested with roles such as data quality manager, data quality analyst, data quality trainer and subject matter expert—provides underlying standards with respect to various dimensions of data quality, defines mechanisms for communicating business uses of data on an ongoing basis, and delineates procedures for evaluating the quality of data. By providing a roadmap for interpreting (metadata) and assessing data, data quality decisions are pivotal in the effective governance of data assets.

Metadata. Defined as "data about data," metadata describes what the data is about and provides a mechanism for a concise and consistent description of the representation of data, thereby helping interpret the meaning or "semantics" of data. Different types of metadata such as physical, domain-independent, domain-specific, and user metadata[8] play a role in the discovery, retrieval, collation and analysis of data. At the lowest level, *physical metadata* includes information about the physical storage of data. *Domain-independent metadata* includes descriptions such as the creator/modifier of data and authorization/audit/lineage information related to the data. By providing a set of mappings from a representation language to agreed-upon concepts in the real world, domain-specific metadata connects a database to the "real world." Domain-specific metadata, for example, can be specified at different levels—such as division and organization; at the division-level it provides descriptions of the application data for individual units, while at the organization-level it supports reconciliation of domain-specific (data) descriptions for the entire organization. Finally, *user metadata* includes annotations that users may associate with data items or collections; such annotations can, for example, capture user preferences and usage history.

The metadata that is employed in

c http://www.dw-institute.com/research/display.aspx?ID=6626

an enterprise depends on the intended use of and access to the data, as well as the management of its life cycle. To support retrieval and analysis of data, the metadata decision domain may be vested in such roles as enterprise data architects and data modeling engineers to develop a programmatic approach for documenting the semantics of data. To ensure that the data is interpretable, standardizing metadata provides the ability to effectively use and track information. As the environment for a business changes, the way an organization conducts business – and consequently the associated data – also changes. As such, there is a need to manage changes in metadata as well.

Data Access. Data access is premised on the ability of data beneficiaries to assign a value to different categories of data. Effective risk analysis by data security officers, for example, identifies the data needs of the business and addresses safeguards to ensure the confidentiality, integrity and availability of data. By integrating risk assessment with an organization's legal and regulatory compliance monitoring efforts (such as requirements of the Graham-Leach Bliley Act for financial industry), industry standards serve as a guide for the writing and updating of an organization's access policies and standards. The data access standards (and the associated service level agreements) can be based on the definition of "unacceptable" uses of data and external requirements for auditability (the ability to track who/what has accessed/modified data), privacy and availability. Data access decisions also provide standards at the physical and logical level.[6] The standards for physical data integrity ensure that the data is immune to physical harm such as power failure; the standards for logical data integrity ensure that the structure of a database is preserved. Developing integrated, enterprise-wide data access decisions can also help automate the migration of data from over-utilized volumes into under-utilized volumes across DAS/NAS/SAN environments.

Data Life cycle. Realizing that all data moves through life-cycle stages is central to designing data governance. From the perspective of data in an electronic health record (EHR) maintained by a hospital, the uses and thereby the value of the diagnostic information of a patient admitted in the hospital changes as the patient undertakes surgery, moves to an acute care center, is discharged, receives a follow-up consultation, and transitions from sick-care to wellness-care. By understanding how data is used, and how long it must be retained, organizations can develop approaches to map usage patterns to the optimal storage media, thereby minimizing the total cost of storing data over its life cycle.

Many organizations do not know what data they have, how critical that data is, the sources that exist for critical data, or the degree of redundancy of their data assets.[4] In order to manage the inventory of data as well as its various data sources, information chain managers[d] develop an understanding of different types of data that are the most/least prevalent, their storage requirements, and the growth trends. A data taxonomy can help in the management of the lifecycle of data, which in turn can be embedded as metadata; additionally, service level agreements (for data access/use) can also be embedded as metadata. By placing data on an appropriate storage medium according to business needs, data can be more effectively distributed across multiple resources, thus leading to improved storage utilization and reduced storage acquisition costs.

Besides cost imperatives, compliance issues related to legislation, such as HIPAA, SOX and Basel II, determine how organizations must deal with the lifecycle of data, its retention and archival. Archive and backup are not synonymous. When a file is archived, it is usually deleted from the source and replaced with a metadata pointer that enables its retrieval from the archive; additionally, the archive is usually indexed. In contrast, a backup involves saving a large block of (snapshot) data on a secondary storage medium, which provides temporary protection of data.

Table 2: Potential example of data governance matrix.

Locus of accountability \ Decision Domain	Data Principles	Data Quality	Metadata	Data Access	Data Lifecycle
Centralized	✓				
↕				✓	✓
			✓		
↕		✓			
Decentralized					

Assessing Data Governance

To design data governance, we have presented an overall framework that provides a set of five data decision domains. By specifying data decision domains that are consonant with IT decision domains, we have also provided an overarching framework to align the IT assets with the data assets (see Table 1). IT infrastructure includes decisions that determine shared and enabling services and the capabilities to enable tracking, storing, analyzing, modeling and presenting data. As may be evident, the decisions related to IT governance are related to those for data governance; similarly, data governance decisions should be tightly integrated with those in IT governance. As such, defining common mechanisms across data and IT assets could induce improved performance. For example, the same committee that establishes the role of IT in business (IT principles) could be employed to clarify the role of data as an asset (data principles).

In designing data governance, the assignment of the *locus of accountability for each decision domain* will be somewhere on a continuum between centralized and decentralized.[2] Table 2 provides an example of what a data governance matrix, which includes locus of data decision making accountability for each of the five decision domains, could be for a given organization. For example, the decision rights for defining the organization's data principles could be highly centralized within a group of corporate

d https://www.research.ibm.com/journal/sj/464/vayghan.pdf

executives who serve as data trustees. In contrast, decision rights for data quality may belong to business managers who are data owners in many different business units, and thus be highly decentralized. Decisions related to data access and data life cycle may be vested with an enterprise data architect and a data security officer, respectively, as the hub, but with business unit participation but not authority (such as, data beneficiaries) as the spokes. Finally, the decision rights for the metadata domain may involve both data consumer and data modeling engineers, and a more balanced approach to responsibility and accountability; hence, it is modeled here as at the midpoint on the continuum.

Both structural and non-structural mechanisms[2] can be employed to implement the governance structure shown in Table 2. For example, a committee of business leaders may review and approve IT project requests and/or act as the governing body for developing and enforcing a set of data principles. For other decision domains that require collaboration across business unit and IS professionals, similar standing committee mechanisms can also be employed, as well as processes that help ensure consistent behaviors across multiple business and IS units. Corporate announcements and other central communications using Web-based portals could be the mechanisms employed to disseminate policy decisions and procedures, as well as to convey the organization's data governance objectives. Finally, organizational incentives and reward systems could be designed to reinforce the value that the organization places on managing data as an organizational asset.

Conclusion

We have presented a data governance framework that can be used by practitioners to develop a data governance strategy and approach for managing data as an organizational asset. We have identified five decision domains, presented arguments for why each of these domains is important, described some key decisions to be made for each domain, and provided some examples of organizational positions that may be given accountability.

We also have proposed that differing levels of centralized, decentralized, and shared decision rights may be appropriate for different decision domains in the same organization. Similar to Weill and Ross,[10] we also suggest that a "one page" design matrix (Table 2) may be useful for communicating a given organization's data governance approach. The proposed framework also provides a common terminology that can be used by researchers to share their findings with other members of the IS community. C

References

1. Ballou, D. P. and Pazer, H. L. Modeling data and process quality in multi-input, multi-output information systems. *Management Science 31*,(1985), 150-162.
2. Brown, C. V. Horizontal mechanisms under differing IS organization contexts. *MIS Quarterly 23*, (1999), 421-454.
3. Griffin, A. and Hauser, J. R. The voice of customer. *Marketing Science 12*, (1993), 1-27.
4. Levitin, A. V. and Redman, T. C. Data as resource: Properties, implications, and prescriptions. *Sloan Management Review*,(1998), 89-101.
5. Olson, J. E. *Data Quality: The Accuracy Dimension*. Morgan Kaufmann, San Francisco, CA, 2003.
6. Pfleeger, C. P. and Pfleeger, S. L. *Security in computing*. Prentice Hall, Upper Saddle River, NJ, 2003.
7. Redman, T. C. The impact of poor data quality on the typical enterprise. *Comm. ACM 41*, (1998), 79-82.
8. Singh, G., Bharathi, S., Chervenak, A., Deelman, E., Kesselman, C., Manohar, M., Patil, S., and Pearlman, L. A metadata catalog service for data intensive applications. In *Proceedings of the ACM/IEEE SC2003 Conference on High Performance Networking and Computing*. (Phoenix, AZ, 2003)
9. Wang, R. Y. and Strong, D. M. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems 12*, (1996), 5-34.
10. Weill, P. and Ross, J. W. *IT governance: How top performers manage IT decision rights for superior results*. Harvard Business School Press, Boston, MA, 2004.

Vijay Khatri (vkhatri@indiana.edu) is an associate professor at the Kelley School of Business of Indiana University, Bloomington, Indiana, USA.

Carol V. Brown (Carol.Brown@stevens.edu) is a distinguished professor at the Howe School of Technology Management of Stevens Institute of Technology, Hoboken, New Jersey, USA.