Operationalizing Data Governance via Multi-level Metadata Management

Stefhan van Helvoirt and Hans Weigand(✉)

Tilburg School of Economics and Management, Tilburg, The Netherlands
svhelvoirt@gmail.com, h.weigand@tilburguniversity.edu

Abstract. Today's rapidly changing and highly regulated business environments demand that organizations are agile in their decision making and data handling. At the same time, transparency in the decision making processes and in how they are adjusted is of critical importance as well. Our research focusses on obtaining transparency by not only documenting but also enforcing data governance policies and their resultant business and data rules by using a multi-level metadata approach. The multi-level approach makes a separation between different concerns: policy formulation, rule specification and enforcement. This separation does not only give more agility but also allows many different implementation architectures. The main types are described and evaluated.

Keywords: Data warehouses · Data governance · Metadata · Business rule enforcement

1 Introduction

The amount of data that is available in the digital universe is growing at an exponential rate and will only continue to grow with the rise of new technologies such as the Internet of Things. Nowadays data is more important than ever before due to the speed of business change. This is emphasized with the rise and use of Master Data Management (MDM) systems in the last decade. MDM adds a new dimension to the data that focusses on establishing integration and interoperability of heterogeneous databases and applications in a business oriented manner [1, 2].

Recent studies have shown that organizations that are capable of effectively utilizing and analyzing their data outperform their competitors [5]. In order to actively use the data that is available both within and outside the organization, the organization must find a way to actively and sufficiently tag the data with metadata [9]. Especially in a new digital world in which organizations are rapidly integrating data from various heterogeneous sources. This need is emphasized with the rise of new data warehouse platforms such as IBM's Data Reservoir. Having a proper data governance program in place is crucial for effectively managing the data that resides in such aggregated environment [3].

What is important in today's highly regulated business environments is not only effective data governance but also that the governance is transparent and auditable. For instance, exporting and importing shippers need to comply with tax regulations and

M. Janssen et al. (Eds.): I3E 2015, LNCS 9373, pp. 160–172, 2015.
DOI: 10.1007/978-3-319-25013-7_13

customs security controls. It is very hard for companies to prove compliance if the data infrastructure is not well-controlled in a transparent way. Sometimes data governance is mandatory by law as with BASEL BCBS 239, effective from 1/1/2016.

Our research goal is obtaining adaptability and transparency by not only documenting but also enforcing data governance policies and their resultant business and data rules. In this paper, we introduce a multi-level framework and use it to evaluate the current capabilities of IBM's InfoSphere package as used in its Data Reservoir solution, while also providing incentives to further extent the governance capabilities. Additional layers of logic are added to reify governance policies in data movement, applications and databases. A preamble on Data Governance and metadata is provided in Sect. 2, to lay the foundation for our multi-level metadata framework discussed in Sect. 3. Section 4 continues with an overview of various implementation styles to establish a Data Governance environment and Sect. 5 evaluates IBM InfoSphere offerings to establishing operationalized Data Governance.

2 Background

2.1 Data Governance versus Data Management

According to Khatri and Brown [6], based on Weill and Ross, "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." Data management activities focus on the development and execution of architectures, policies, practices and procedures to enhance and manage the information lifecycle within a specific application and mostly during data entry/creation. Data Governance on the other hand also includes aggregated and integrated data that is made available as a data asset within the organization.

The Data Governance domain consists of three focus areas; people, processes and technology. Many publications on data governance focus primarily on the people and processes aspects of implementing a data governance program. We can use IBM's holistic approach to Big Data governance as an example, which consists out of the following six sets: define business problem, obtain executive sponsorship, align teams, understand data risk and value, implement analytical/operational projects and measure results. Its focus has been primarily on the first four steps. Our focus is mainly on the "implement analytical/operational projects" from a technical perspective as this appears to be research gap. However, all steps in the holistic approach are needed in order to have an efficient and reliable data governance program. Capturing and enforcing business rules, without the proper knowledge of the available data, its value and the interdependencies between data is undoable and undesirable. We do not underestimate the political change that is needed to transform the organization into an information-driven environment. An overall transition needs to be made from thinking and developing individual applications to a unified acceptance and usage of data as the foundation of information and knowledge [8].

2.2 Data Quality and Trust

Governance is more than achieving compliance [7]. Achieving data governance has to do with adopting practices and principles that increase data quality and trust. Having established data quality and trust, the organization can start using their data in a reliable and controlled manner and evaluate its data usage and governing capabilities by implementing appropriate metrics. A valuable data quality standard currently in development is ISO 8000. It focusses on data characteristics and exchange in terms of vocabulary, syntax, semantics, encoding, provenance, accuracy and completeness.

2.3 Metadata as Indispensable Enabler

The importance and utilization of metadata has been increasing rapidly over the last decade as metadata is making its transition from a technical aspect to a business necessity. Metadata is needed for establishing data quality and turning data into understandable information that can be consumed by both business/IT users and software for automation.

Looking at publications on metadata form the past fifteen to twenty years shows that there are various types and classifications of metadata, each with its own specific purpose and granularity. This paper uses the metadata framework as presented by Ron Klein (KPMG) at the 2014 ECCMA conference [7]. This metadata framework consists of three vertical levels and three horizontal levels. The vertical levels are *business*, *technical* and *operational*, while the horizontal levels consist of the categories *descriptive*, *administrative* and *lineage*. Business metadata includes business terms, data owners, stewards, and governance policies and business rules governing the data. Technical metadata is used for tool integration to manage, transform and maintain the data. Examples of technical metadata are database system names, table and column names, code values and derivation rules. Operational metadata contains run-time information e.g. last load, usage statistics and log reports. In short, business metadata has a value and meaning for business oriented users, technical metadata is used primarily by Extract, Transform, Load (ETL) developers while operational metadata is used to provide insights in data usage and rule validation.

3 Multi-level Metadata Management

In the following, we will focus on the policies *behind* the meta-data as such, for instance, policies and access control, policies on quality requirements, or policies on the use of semantic standards, rather than the meta-data tags themselves. Our goal is transforming descriptive data governance policies into implementable rules by using a multi-level metadata approach. The multi-level metadata model consists of four levels as depicted in Fig. 1. Distinguishing these levels leads to maximal adaptability and transparency (cf. [4]).

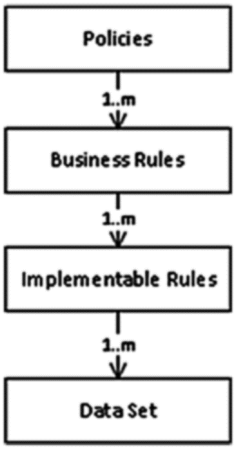

Fig. 1. Multi-level Metadata Management framework

Policies are abstract formulations of business goals, desirable behavior, guidelines and generally accepted practices. Business rules are a formalization of a (partial) aspect of a policy, stating the trigger and follow-up action in a structural natural language consisting of business terminology. Implementable rules are executable objects that contain the logic needed to enforce the business rule. How a business rule is implemented and enforced is entirely dependent on enforcement strategy of that specific rule. As mentioned by Weigand et al. [14], "although business rules are more formal than policies, they are still at the level of business requirements (...), rather than execution. They model "what" is required, rather than "how" it should be implemented" [14]. The same paper emphasizes the need to clarify the enforcement strategy and provides four types of enforcement: preventive, punitive, corrective and adhortative. Here adhortative means that the responsible user is requested to solve the violation when it is detected; the system does not prevent or correct it itself. Punitive means that a sanction is given on violation of the rule whereas corrective means that the system automatically corrects the violation and moves forward to a consistent state, typically by means of compensation. When the enforcement is separated from the business rule specification as such, this allows for great flexibility: the company can switch rather easily from a more loose adhortative approach to a strict preventive approach (or vice versa) depending on the desired compliance levels and operational costs.

Where the enforcement of a business rule takes place is highly dependent on the chosen enforcement strategy and goal of the business rule. For example, a policy stating that all telephone numbers should be formatted according to the applicable standard of the country that it applies to, will have a business rule declaring the use of a given standard for telephone numbers within a given region. This rule could be enforced at the point of data creation using a preventive strategy, or when data is analyzed, transformed and moved to a different location using a corrective strategy.

Business rules are defined as condition action (CA) rules and require a structural transformation to become executable condition action (ECA) rules.

4 Implementation Archetypes

4.1 Business Rule Extraction

For the enforcement of business rules, we start from the generally accepted approach expressed, among others, by Pierre Bonnet in his book on Enterprise Data Governance [2] in which the business knowledge is extracted from the software (hard-coding) and is presented to the business users in an environment that they can (partially) control. "Maintaining knowledge, in particular within complex and evolving organizations that characterize modern companies, cannot survive the trap set out by fixed and stratified hard-coded software, nor informal (textual) documentation, rarely up to date and non-executable". According to Bonnet, a software package must first be able to interact with an MDM system, before it can demonstrate its ability to enforce the relating business rules (BRMS) that will eventually affect the processes (BPM). "First the data, then the rules and finally the processes" [2]. This way, rules are defined per data domain and not based on the software package that uses the data.

4.2 Enforcement Architectures

Isolating data governance rules from the code is one thing, but still leaves many choices on how to enforce the rules. Based on our analysis, we distinguish between a *decentralized*, *centralized* and *leveled* implementation archetype. The archetype that is most applicable to a given situation depends on the available resources and business requirements [15]. For example, an analytical driven environment will have a specific way of enforcing policies as data is collected from various sources and ingested into one or multiple repositories designed and optimized for specific analytical computations (e.g. IBMs Data Reservoir). The enforcement of policies in such an environment could largely occur at the processing of data movement. On the other hand, enforcing governance policies on the actual applications/databases that create/store the data would require a different approach to integrating and enforcing policies. The difference in these three implementation styles as described in this section, is purely in the area of policy *enforcement*. Our base assumption is that policy and asset descriptions are high level and should not be restricted or influenced by the underlying technology and infrastructure. Furthermore, capturing, defining and maintaining the definition and description of policies and assets at domain or organizational level allows for greater consistency, transparency and manageability. However, this integration also has its costs and concerns. One of the concerns is that responsibility for some resource, including data, should not be taken away from the agents owning the data.

We start off by illustrating and defining the decentralized implementation style, displayed in Fig. 2. In this example we have four data storages, each containing the (business) definitions of applicable policies (no pattern) and the resulting implementation code (striped pattern). The decentralized implementation is very common in situations where data governance maturity is low. This implementation style has some benefits and limitations as illustrated in the Table 1.

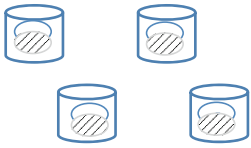

Fig. 2. Decentralized implementation

Table 1. Benefits and limitations of decentralized implementation

Benefits	Limitations
Enforcement of policies as close to the source of data as possible	Siloed knowledge resulting in a lack of reusability and increased risk of inconsistency among separated data storages
Less dependencies and decreased systematic risk	Monitoring compliance and conducting audits is costly and time consuming

Although there are some benefits to mention for the decentralized implementation style, these do not outweigh the limitations. Especially in today's rapidly growing digital ecosystem, in which data is being created by an increased amount of utilities both within and outside the organization. A leveled or centralized implementation style would deliver a more feasible and desirable approach to enforcing data governance, however this requires the presence of a central governance catalog like system for centrally storing and defining data governance policies and assets (cf. [11]).

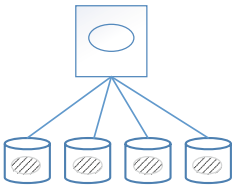

Fig. 3. Leveled implementation

The leveled (Fig. 3) and centralized (Fig. 4) implementation styles make use of a central governance catalog. The leveled implementation style uses a centralized governance catalog repository which is capable of storing the data asset definitions and data governance policies. These asset descriptions describe both the business characteristics of a dataset (business definition using business terminology, owner, steward etc.). Business assets and policies are linked to denote which policies should apply to a specific asset. The implementation and enforcement of these policies is conducted at the System of Record/Reference (SoR). This approach allows for the creation and maintenance of

both asset descriptions and policies at a central level, allowing for greater transparency and consistency. At the same time, the implementation can make use of the tools most efficient for the particular SoR. However, there are also some drawbacks to this implementation style as shown in Table 2. To address the consistency problem, one could imagine an automated update system that pushes any changes in the policy definitions forward to the SoRs. However, when the diverse SoRs use different local enforcement tools, such an update may also require as many compilations as there are different SoRs.

Table 2. Benefits and limitations of leveled implementation

Benefits	Limitations
Consistency in asset and policy definitions	Gap between the definition of a policy and the actual implementation which could result in misinterpretation and incorrect enforcement
Increased transparency in the available data and the rules that shape the data and its use throughout the data lifecycle	Lack of consistency in the enforcement of policies due to high diversity of SoR sources

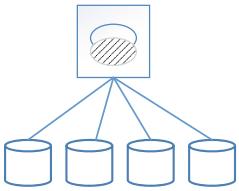

Fig. 4. Centralized implementation

Lastly, we have the centralized implementation style as displayed in Fig. 4. A centralized implementation requires the presence of a governance catalog repository with enhanced and additional capabilities. We can distinguish at least three variants of centralized systems capable of establishing and enforcing governance policies on data assets. These variants differ in how users get access to the data: distributed or intermediated. The first being an environment in which end users direct their requests to the various source applications or databases that in turn call a *service* running on the governance catalog system to evaluate, conduct and if needed enforce a policy. In the second environment, the governance catalog functions as an *intermediary* that ingest data from various sources and uses ETL practices to conduct and enforce governance policies before providing the data to the requesting end user. Lastly is a *hybrid* environment of having both the capabilities of data movement/integration (second environment) and service-like enforcement (first environment).

The hybrid environment contains the highest level of complexity to implement as it requires functionality for two entirely different environments. The first service-based environment excels in a landscape in which policies are defined for enforcing rules designed for the data creation phase (rather than retrieval phase), while the second "intermediary" environment excels in a more analytical landscape in which data needs to be collected, aggregated and delivered to an end user or analytical application (e.g. SPSS). In the first environment we described, the governance catalogs functions primarily as a rule engine. In the second environment the governance catalog functions like a true catalog that controls the data flows from source systems to end users based on the governance policies that are defined. Both environments are needed to establish a holistic data governance solution. Table 3 summarizes the benefits and limitations of a centralized implementation strategy.

Table 3. - Benefits and limitations of centralized implementation

Benefits	Limitations
Optimal consistency in asset and policy definitions and enforcement	May lead to higher network load, possibly lower enforcement efficiency, Single Point of Failure
Allows full integration of policies	Requires high level of integration (organizational and technical)

Technological advances and the use of Enterprise Application Integration (EAI) and Service Oriented Architecture (SOA) to develop new applications and services help in establishing this service-oriented environment for discovering and analyzing of data. EAI allows for the extraction of business policies and rules from the applications, creating increased flexibility and agility. SOA is a framework to "address the requirements of loosely coupled standards-based and protocol-independent distributed computing, mapping enterprise information systems appropriately to the overall business process flow" [10]. Technological advances include, amongst others, new ways of processing data (e.g. NoSQL, in-memory, Hadoop), a decrease in storage costs and increase in memory and computing power to perform the needed operations and a move to semantic systems. SOA can be enhanced with semantic technologies, for instance, to improve service identification [13].

4.3 Catalogue Virtualization

Although a centralized governance catalogue has important management advantages, the drawback is that business users – in particular, the managers responsible for the data –are set on a distance. This can be remedied by virtualizing the catalogue. This means that the various data policies are stored in a distributed way, under the control of the business user. These business users are at various levels: company-wide standards are maintained at corporate level, other policies at division of department level. In the simplest form of virtualization, these distributed data

policies are just synchronized regularly with the central governance catalog. Alternatively, there is only a virtual central catalog, the combination of all distributed policies. In both cases, we assume that policy owners receive feedback (dashboard) on the actual policy compliance.

A critical issue in such a solution is the consistency of the policies. Policies may be conflicting. For instance, a corporate policy may be that all management reports are readable for the internal audit group, whereas a manager may want to restrict access to members of his own department only. One business user may want to express weights in kg and another one in pounds. In the context of this paper, we just mention a few alternative solutions which roughly correspond to the general rule enforcement strategies that we mentioned in Sect. 3. One is to accept inconsistencies as a fact of life and include meta-rules for solving them. A meta-rule can be based on the company hierarchy where corporate policies overrule local ones. This corresponds to a corrective approach because it effectively makes changes in the policies – not in their formulation, but in their application. Alternatively, we can take an adhortative approach that accepts inconsistencies but stimulates policy owners to avoid them at specification time. Closely related, a lazy evaluation (corresponding to a detective approach) can be used that detects conflicts when they actually occur and reports them back to the policy owners. This can be a pragmatic approach in situations of relative low governance where the probabilities of actual conflicts are low. Finally, the most rigid approach is to prevent any inconsistency by using a consistency checker before any policy is deployed. This is a challenge in a distributed environment, although in principle, such a checker is not different from the checkers in a centralized catalog. Last but not least, it is not necessary to choose only one approach. For instance, the company may use a preventive approach for all data standard policies and a detective approach for data access policies.

Once a virtual solution is in place, a next step can be to relax the centralization of the governance catalog. In large companies, a completely centralized approach is not realistic. Some distribution in “regions” or “zones” is unavoidable. In such a situation, a business user may be connected to one region, but also with more regions. Locally, he can manage his policies for both. Data traffic between regions is based on agreements that appear as policies in each of the regions involved.

The virtual solution described in this section can be combined smoothly with a strict distinction between “policy” and “rule” level, as sketched in Sect. 3. This means that the business users publish policies in a user-friendly policy language that is translated to formal business rules on the central catalog (physical or virtual).

5 State-of-the-Art Governance Solutions: IBM InfoSphere

In this section, we analyze in depth one commercially available solution in data governance, IBM InfoSphere. Since this is considered state-of-the-art technology, it can be seen as representative. Our goal for this study was not to compare it with other products, but to see to what extent a multi-level governance model is or can be implemented with this solution. Our analysis is based on the system documentation, expert interviews, and user experience.

5.1 Description

Our evaluation of the IBM InfoSphere suites capability to define and enforce governance policies focuses on IBM InfoSphere Information Governance Catalog (formerly known as InfoSphere Business Information Exchange) and IBM InfoSphere DataStage. Additional tools such as IBM InfoSphere Information Analyzer, IBM InfoSphere Optim and IBM InfoSphere Guardium are used to illustrate specific enforcement examples. Information Governance Catalog is designed to contain both the business glossary (terminology) as well as a list of all available information assets (e.g. dataset, table, policies, and rules) and a variety of additional metadata to describe and define the asset. An information asset is defined as "a body of information, defined and managed as a single unit, so that it can be understood, shared, protected and exploited effectively. Information assets have recognizable and manageable value, risk, content and lifecycles" [12]. Information Governance Catalog allows for the creation of a hierarchical structure to define the relations between policies and rules. These rules can be assigned to a business term. The business term defines and references to the actual source of the authoritative data.

Enforcing data governance policies focusses primarily on achieving a compliance layer. The compliance layer consists of four areas; Policy Administration, Policy Implementation, Policy Enforcement and Policy Monitoring. A *policy* is a (natural language) description of business intent for a class of assets to adhere to a

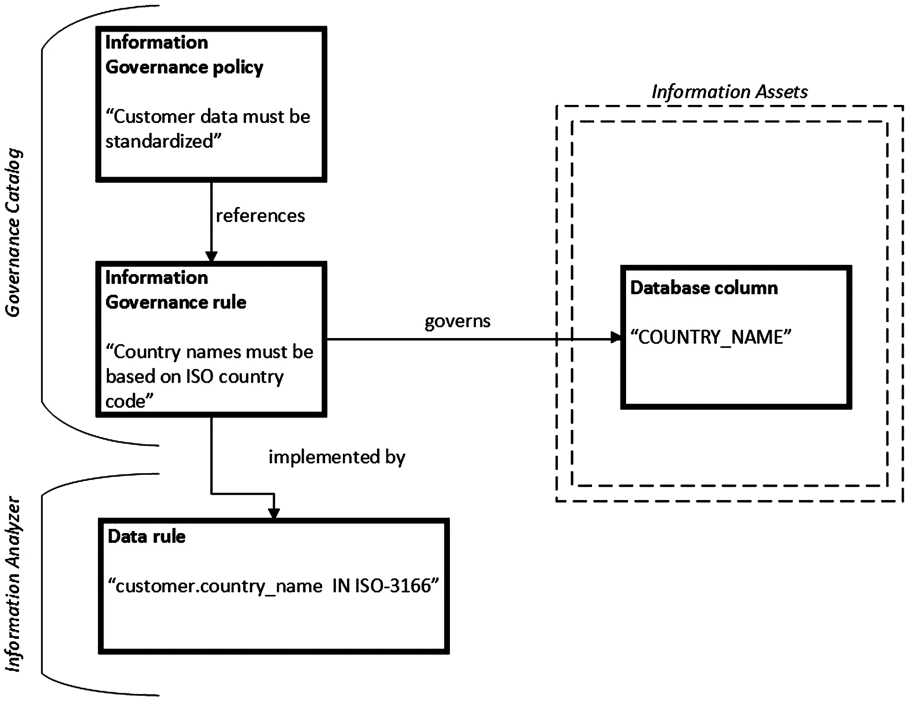

Fig. 5. Infosphere information governance

certain behavior. A *rule* defines how a policy will be implemented, it contains the policy response. Additionally to policies and rules are the *control* and *enforcement points*. A control point is a collection point for evidence that a policy is being complied with. An enforcement point is where a policy implementation (rule implementation or process) is executing. Enforcement points can be seen as hard enforcement measures that guarantee preventive compliance. Control points provide the soft enforcement which is used in the remediation that occurs after the fact. Soft enforcements are used in policies defined at a more abstract level that cannot be hard enforced.

Figure 5 illustrates how information governance policies, governance rules, data rules and the information assets are related. There is some correspondence to the multi-level meta-data framework described above. One difference is that the information assets are related to the information governance rules (business rules), rather than implementable data rules, but if there is 1–1 relationship between data rule and information governance rule, the two representations are equivalent. However, there is no formal representation for the information governance policies and rules, and hence it is not possible to check the consistency of the rules or adapt them automatically.

5.2 Evaluation

Does the IBM InfoSphere suite support multi-level metadata to enable governance, and to what extent? To answer this question, we have looked at the capabilities to validate and enforce rules in the process of moving data (ETL) by using primarily IBM InfoSphere DataStage. Within DataStage we have the capability to create jobs for performing various ETL activities. These jobs can be assigned to a rule in the Governance Catalog as the implementable artifact. Executing a DataStage job results in the creation of operational metadata which is used to establish lineage and provide metadata to the governance dashboard. The operational metadata is mapped to IBM's private proprietary metadata model called XMeta. Besides generating data lineage the capabilities include measuring data quality and values using IBM InfoSphere Information Analyzer. Having insights into the quality and usage of data creates a tremendous increase in transparency for both business and technical users.

The Information Governance Catalog should be used as the central storage point for all the metadata that is needed for providing sufficient insights in definitions, usage, accountability and compliance. However, the current capabilities of enforcing rules and measuring their results requires a lot of technical expertise (ETL development etc.), which is undesirable in an environment that should be business driven. A more formal (semantic) approach to defining the rules should empower the business users with more capabilities to governing "their" data. Policy and rule administration are currently defined in free-text format, which could result in misinterpretations during the implementation and enforcement phase, and creates an opaque environment. A BRMS that utilizes the capabilities of defining rules in a natural structured language reduces opaque and misinterpretation, resulting in a more transparent environment. Answering the research question: IBM InfoSphere suite supports multi-level metadata to enable governance, but there is still a lot of room to enhance these

capabilities to increase transparency and formalization. At the moment, it supports typically a leveled approach, not full centralization.

6 Conclusion

In order to operationalize data governance, the implementing organization needs to have the resources and capabilities in place to define and enforce data governance policies and rules. Using a multi-level metadata framework we created an insightful segregation between defining the policy and rule specifications and the resultant implementation of rules and jobs. With this segregation in place, and the capability of empowering qualified business users to define governance specifications, allows for better adaptability and transparency of data governance. As this paper presented, there are various ways of implementing a multi-level metadata framework. Having the capability to enforce policies and rules both in a centralized and decentralized manner, allows for the most flexibility. However, specific software might need to be purchased to establish an environment for this in the form of a governance catalog. IBM InfoSphere suite provides most of the capabilities needed to start operationalizing a multi-level data governance program, but formalization of policies and rules is needed in order to get to a higher level, in particular, to one supporting self-adaptation.

References

1. Baca, M., Gill, T., Gilliland, A.J., Whalen, M., Woodley, M.S.: Introduction to Metadata, Revised edn. Getty Publications, Los Angeles (2008)
2. Bonnet, P.: Enterprise Data Governance: Reference and Master Data Management Semantic Modeling. Wiley, New York (2013)
3. Cheong, L., Chang, V.: The need for data governance: a case study. In: Toowoomba: 18th Australasian Conference on Information System (2007)
4. Gong, Y., Janssen, M.: From policy implementation to business process management: principles for creating flexibility and agility. Gov. Inf. Q. **29**, S61–S71 (2012)
5. IBM Center for Applied Insights. Outperforming in a data-rich hyper-connected world. New Orchard Road: IBM Corporation (2012)
6. Khatri, V., Brown, C.V.: Designing data governance. Commun. ACM **53**(1), 148–152 (2010)
7. Klein, R.: Metadata is 'not' a technical term anymore: frame to work. In: 2014 International Data Quality Summit, ECCMA 2014: KPMG, pp. 1–18 (2014)
8. Marco, D.: Practical steps for overcoming political challenges in data governance. In: IDQSummit 2014 ECCMA 2014: EWSolutions, pp. 1–37 (2014)
9. NISO. Understanding Metadata. National Information Standards Organization (2004)
10. Papazoglou, M., van den Heuvel, W.-J.: Service oriented architectures: approaches, technologies and research issues. VLDB J. **16**(3), 389–415 (2007)
11. Singh, G., Bharathi, S., Chervenak, A., Deelman, E., Kesselman, C., Manohar, M., .Pearlman, L.: A metadata catalog service for data intensive applications. In: Supercomputing ACM/IEEE Conference (2003)
12. The National Archives. The Role of the Information Asset Owner: a Practical Guide. National Archives (2010)

13. Vitvar, T., Peristeras, V., Tarabanis, K.: Semantic Technologies for E-Government: an Overview. Springer, Berlin (2010)
14. Weigand, H., van den Heuvel, W., Hiel, M.: Business policy compliance in service-oriented systems. Inf. Syst. **36**(4), 791–807 (2011)
15. Wende, K., Otto, B.: A Contingency Approach to Data Governance. In: MIT Information Quality: ICIQ (2007)