Fall 2011

**Effecting Data Quality Through Data Governance: a Case Study in the Financial Services Industry**

Patrick Egan  
Regis University

Follow this and additional works at: [http://epublications.regis.edu/theses](http://epublications.regis.edu/theses)

Part of the [Computer Sciences Commons](http://epublications.regis.edu/theses)

**Recommended Citation**

Use of the materials available in the Regis University Thesis Collection ("Collection") is limited and restricted to those users who agree to comply with the following terms of use. Regis University reserves the right to deny access to the Collection to any person who violates these terms of use or who seeks to or does alter, avoid or supersede the functional conditions, restrictions and limitations of the Collection.

The site may be used only for lawful purposes. The user is solely responsible for knowing and adhering to any and all applicable laws, rules, and regulations relating or pertaining to use of the Collection.

All content in this Collection is owned by and subject to the exclusive control of Regis University and the authors of the materials. It is available only for research purposes and may not be used in violation of copyright laws or for unlawful purposes. The materials may not be downloaded in whole or in part without permission of the copyright holder or as otherwise authorized in the “fair use” standards of the U.S. copyright laws and regulations.
EFFECTING DATA QUALITY THROUGH DATA GOVERNANCE: A CASE STUDY
IN THE FINANCIAL SERVICES INDUSTRY

A THESIS

SUBMITTED ON 26TH OF AUGUST, 2011

TO THE DEPARTMENT OF INFORMATION SYSTEMS
OF THE SCHOOL OF COMPUTER & INFORMATION SCIENCES
OF REGIS UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF MASTER OF SCIENCE IN
SOFTWARE ENGINEERING AND DATABASE TECHNOLOGIES

BY

Patrick Egan

APPROVALS

Ernest Eugster, Ph.D.

Donald J Ina - Faculty of Record

Nancy Berkhemer

Program Coordinator
Abstract

One of the most significant challenges faced by senior management today is implementing a data governance program to ensure that data is an asset to an organization’s mission. New legislation aligned with continual regulatory oversight, increasing data volume growth, and the desire to improve data quality for decision making are driving forces behind data governance initiatives. Data governance involves reshaping existing processes and the way people view data along with the information technology required to create a consistent, secure and defined processes for handling the quality of an organization’s data. In examining attempts to move towards making data an asset in organizations, the term “data governance” helps to conceptualize the break with existing ad hoc, “siloed” and improper data management practices. This research considers a case study of a large financial services company to examine data governance policies and procedures. It seeks to bring some information to bare on the drivers of data governance, the processes to ensure data quality, the technologies and people involved to aid in the processes as well as the use of data governance in decision making. This research also addresses some core questions surrounding data governance, such as the viability of a “golden source” record, ownership and responsibilities for data, and the optimum placement of a data governance department. The findings will provide a model for financial services companies hoping to take the initial steps towards better data quality and ultimately a data governance capability.
Acknowledgements

To my wife Janice and kids Brendan and Kevin – Thank you for your support, understanding and sacrifice for the last two years.

To my parents Paddy and Bridget for giving me the encouragement to go back to school after all these years.

To my thesis supervisor Dr. Ernest Eugster, for his advice and guidance throughout the development of this thesis.

To Pioneer Investments IT staff for giving me the opportunity to work with some of the best minds and technologies in the financial industry.

To the staff and students of both NUI Galway and Regis University’s MScSIS program 2010 – 2011 for giving me the opportunity to learn.
Table of Contents

Chapter 1 Introduction .................................................................................................................. 1

1.1 Where Does Data Governance (DG) Come From? .............................................................. 2

1.2 Thesis Statement .................................................................................................................... 6

1.3 Research Methodology ........................................................................................................ 6

1.4 Success Criteria .................................................................................................................... 7

1.5 Summary ............................................................................................................................... 8

Chapter 2 Literature Review ......................................................................................................... 9

2.1 The High Cost of Bad Data .................................................................................................. 9

2.1.1 Sources of Bad Data ...................................................................................................... 12

2.2 Data Governance Best Practices ....................................................................................... 14

2.2.1 Data Governance Maturity Model ............................................................................... 16

2.3 Master Data Management .................................................................................................. 19

2.3.1 MDM Definition ........................................................................................................... 19

2.3.2 Benefits of Master Data Management ........................................................................... 22

2.3.3 MDM Structure ............................................................................................................. 26

2.3.4 Registry Hub Patterns ................................................................................................... 27

2.3.5 Coexistence Hub Patterns ............................................................................................ 28

2.3.6 Transactional Hub Patterns .......................................................................................... 29
4.2 Organizational Structure .................................................................................. 78

4.3 Data Quality ........................................................................................................ 80

4.3.1 Data Management Framework (DMF) .............................................................. 82

4.3.2 Quality Summary .............................................................................................. 87

4.4 Metadata Management ....................................................................................... 87

4.4.1 Business Glossary ............................................................................................ 88

4.4.2 Metadata Summary ........................................................................................... 93

4.5 Master Data Management .................................................................................. 93

4.5.1 MDM Summary ............................................................................................... 97

4.6 Survey .................................................................................................................. 97

4.6.1 What is the Status of Your DG initiative? .......................................................... 98

4.6.2 Data Governance Competencies ....................................................................... 98

4.6.3 Data Governance is a Function of which Group(s) ............................................. 99

4.6.4 Data Governance Drivers ................................................................................ 100

4.6.5 What is the Scope of the DG initiative? ............................................................. 100

4.6.6 Obstacles to Data Governance ......................................................................... 101

4.6.7 Executive Participation in a DG program should be Positioned ....................... 102

4.6.8 Data Governance Planning .............................................................................. 102

4.6.9 Data Governance Structures .......................................................................... 103

4.6.10 Data Governance’s Importance to the Industry ............................................ 104
4.7 Summary ................................................................. 105

Chapter 5 Conclusions ............................................................. 107

  5.1 Contributions ................................................................. 107

  5.2 Future research .............................................................. 109

Appendix A Survey Form ............................................................ 111

Appendix B Table of Abbreviations ............................................. 115

References .................................................................................. 116
List of Figures

Figure 2.1 Impact of poor data ........................................................................................................... 11
Figure 2.2 1-10-100 Rule increasing cost impact of delayed DQ resolution ...................................... 12
Figure 2.3 Data governance layers ................................................................................................... 15
Figure 2.4 Data governance maturity model ..................................................................................... 17
Figure 2.5 Elements of effective data governance .......................................................................... 19
Figure 2.6 MDM registry hub .......................................................................................................... 28
Figure 2.7 MDM coexistence hub .................................................................................................... 29
Figure 2.8 Transactional hub ............................................................................................................. 30
Figure 2.9 MDM hub complexity ..................................................................................................... 32
Figure 2.10 ETL process and data .................................................................................................... 38
Figure 2.11 Top data integration challenges ...................................................................................... 39
Figure 2.12 Data federation overview ............................................................................................... 45
Figure 2.13 TDQM quality cycle ...................................................................................................... 50
Figure 3.1 Elements of a research project .......................................................................................... 72
Figure 4.1 Higher order organizational structure for IT department .................................................. 79
Figure 4.2 Expanded organizational chart for data governance .......................................................... 80
Figure 4.3 New DQ rule workflow .................................................................................................... 84
Figure 4.4 DMF exception integration ............................................................................................... 86
Figure 4.5 Business Term and its underlying metadata ..................................................................... 89
Figure 4.6 Business glossary lineage ................................................................................................. 91
Figure 4.7 Business term impact analysis ......................................................................................... 92
Figure 4.8 Pioneer's MDM overview ................................................................................................. 96
List of Tables

Table 2.1 Data governance RACI chart ...........................................................................35

Table 3.1 IT resources interviewed .............................................................................74

Table 4.1 Exception cost analysis .................................................................................87
"In God we trust; all others must bring data."

— W. Edwards Deming (Hanks, 2008)
Chapter 1 Introduction

Information is now seen by many as one of the core enterprise assets of an organization. In the current knowledge-based economy, financial organizations measure daily trading in nanoseconds and security positions in the millions. Having the right data at the right time to make the right decision is not a benefit, but crucial component of an organization’s success or failure. Renowned management consultant Peter F. Drucker (1996) pointed to this fact in his writings, referring to information as one of the most critical business differentiators, and depicted it as “society’s” organ for the creation of wealth, and as a facilitator for the conversion of business costs into yields. With such high accolades, one could easily make the assumption that data and information would therefore be prized, managed and leveraged by all financial organizations to gain maximum competitive advantage. Yet in many cases the value and return on data has been overlooked in the rush to deploy faster, quicker and cheaper hardware and software systems to the detriment to the very asset that they store and facilitate, and the one the provides the most business value….corporate data.

The oversight, management and value of information to an organization is not a new phenomena. For many decades information has been seen as the sole responsibility of the information technology (IT) department. Historically, IT managed data centrally, often siloed with limited accessibility and flexibility. This tight control of information facilitated a simple governance program where all data could be identified, validated, secured and managed within the confines of a single department. In stark contrast, today’s information ecosystem have data collections from internal and external providers, integrated through multiple systems, with stores of data now numbered in the petabytes with some information possible located in the cloud.
When one includes the current era of Facebook, Twitter, LinkedIn and other social communities, what one perceives and classifies as information and how it’s collected, harvested and integrated varies dramatically from the perceptions and approaches of even a decade ago. The seemingly one constant in all of this is the ever growing need by the financial services industry for more accurate, integrated and timely data. With volume, variation, regulatory and integrity constraints that the modern financial services organization must operate under, one begins to understand the complexities associated with the measurement, management and governance of information in the modern age. The research in this paper intends to prove that the implementation of an adaptable, operational and maintainable data governance program has become a necessity for today’s financial services institutions.

1.1 Where Does Data Governance (DG) Come From?

Today’s financial services institutions face growing volumes of digital data. Market researcher IDC predicting an annual compound growth of data of 60% (IDC, 2008). The breakdown of this information found that 80-85% of this data is unstructured or semi-structured, while 30-40% of all information stored is no longer of use to the organization (DataFrameworks, 2009). At current trends, the amount of electronic data in existence by the end of 2011 will be ten times the amount that existed in 2006. The ability to add “cheap” disk space compounds this issue by facilitating the hoarding process, thus increasing high end costs such as performance latency, backup and restores and the ability to search and find information. IBM (2011) stated that most medium sized companies spend USD$ 5.3 million just searching their own data. Managing such a large, varied and multi-sourced content proves to be a daunting and complex proposition and one that is prone to error. According to market researcher Gartner (2007), 25%
of critical data in all Fortune 1000 companies is flawed, where information is inaccurate, incomplete or duplicated multiple times. Additionally, a Lodestar (2009) survey found that 84% of financial services companies could not produce a Key Performance Indicator (KPI) around the quality of their data, while incredulously 37% were not even convinced of organizational value of a data quality program.

Despite questionable data quality, financial services companies rely on this information to make their most critical decisions relating to security trades, mortgage offerings and other business actions. Corporate data purchased from external vendors also comes with its hazards as evidenced by the mortgage meltdown of 2009. Aided in part by many of the ratings agencies that provided many of the dubious mortgage security vehicles with inflated ratings and underestimated risks and impacts of defaults, many financial institutions such as AIG, Citibank and Bank of America were placed in such financial jeopardy that they necessitated a government bailout to guarantee their survival. Although the lack of data integrity checks was not the sole reason for the financial crisis, the Casualty Actuarial Society (Francis & Prevosto, 2010) reported that data quality deficiencies did play a significant role in mispricing instruments and providing business intelligence errors that contributed to the meltdown.

Many of the early drivers for data governance within financial services usually came as directives from the compliance, risk and audit departments. These “demands” were perceived by other groups as a necessary overhead to meet regulatory and audit commitments, with little or no intrinsic value to the organization as a whole. In recent years there has been large increase in regulatory oversight as a result of prior corporate transgressions. New legislation enacted includes Gramm-Leach-Bliley, Sarbanes-Oxley, and the newly enacted Dodd-Frank Wall Street Reform act. For multinational organizations, a myriad of additional regulations exist such as EU
backed Basel II and Solvency II proposals. In many cases, significant overlap exists between the different legislative policies allowing organizations with a standardized and systematic approach to information management the capability to meet all regulatory requirements effectively and efficiently.

Despite its early roots in regulatory compliance, data quality management began to evolve, initially taking its lead from the manufacturing quality assurance teachings of Deming and Conway (Conway, 1992). Many of the keys concepts introduced by Deming and Juran such as preventative maintenance, process assessment and continual improvement are also found to be valid practices for data governance. Organizations began to realize the impact the “garbage in, garbage out” approach has on their costly resources, with time spent on non-productive tasks such as searching or trying to fix data errors after the fact. In 2003, a TDWI report estimated the direct cost of bad data at over USD $600 billion, yet this number did not take into account the intangible costs such as missed opportunities, goodwill and resource and corporate morale (Loshin, 2010)

The advent of data warehouses, marts and business intelligence systems increases the need for clean and accurate data to provide the rich analytics and predicative analysis one expects from such systems. With the underlying need for better management of data, many financial institutions have incorporated some degree of quality management into their organization. In many cases the initiative may not be officially labeled as data governance, and actions may be limited in scope to certain departments (such as IT), but the understanding is there, that data is complex and it needs to be managed, validated and integrated in a manner that provides the best business benefit for the company.
Over the years, many definitions for what data governance is and what it entails have been promoted. In some cases, the term has been misused to define specific application technologies (such as e-discovery), project initiatives (data cleanups) and even IT operations. Yet, in its truest form data governance represents none of these, but more accurately reflects an iterative program of continual data assessment and improvement. At its heart, DG provides the “processes, policies, standards, organization, and technologies required to manage and ensure the availability, accessibility, quality, consistency, auditability, and security of data in a company or institution” (Informatica, 2006).

The definition, though important for what it includes, is also equally as important for what it omits – the function and responsibility of the IT department. The data governance program has to be a function of the enterprise to be successful and not left to IT to shoulder the burden of maintenance and accountability. Current trends in data governance are beginning to recognize this fact and have placed many of the management functions outside of IT and into the Line of Business (LOB) function (Karel, 2007). Karel saw the responsibility for data governance as a shared program with the business taking tasks of ownership and integrity while IT continues to be responsible for the technical aspects of data such as storage, integration, delivery and technical implementation of business rules. This continuous effort from all parties lends itself to a coordinated, collaborative and managed use of the information asset that a data governance program lends itself to.

Financial organizations in recent years have become aware of the impact that bad data has on their bottom line. But are they defenseless against the ill effects of dirty data? Unequivocally, companies have the ability to rectify many of the ills that affect their data consumption. Some have initiated “once off” projects to clean up “dirty data.” yet some market
analysis (Teradata, 2010) suggested that without the overarching framework of standards, policy and data lifecycle management, their impact has been low and fail to resolve the root cause of subpar data quality. The question then arises: does the financial services industry see the need for an ongoing data governance program and do they have the commitment, resources and means to plan and operate a comprehensive program? This research aims to prove that data governance is not a nicety within the financial services industry, but a program that is necessary to manage their information, and that this capability is feasible and available to manage the data lifecycle from its initial inception to its ultimate disposition.

1.2 Thesis Statement

The exponential growth of data, aligned with its pervasive nature into every facet of a financial services organizational domain necessitates the implementation of robust data governance program to manage data’s lifecycle from its inception to its ultimate disposal.

1.3 Research Methodology

To test this thesis, this researcher adopted a three phase methodology approach. Phase one examined the existing literature, in both general IT domain and within the data governance function. The literature review examined existing academic research and papers in the fields of data quality, master data management and information governance. The information gathered during the literature review provided the general conceptual model for data governance based on academic and commercial best practices.

Phase two examined the implementation and ongoing operations of a data governance program within a major financial services company. The case study included interviews with the
resources leveraged during the implementation and the ongoing support of the program. This analysis provided a practical, real world example of a data governance program in action and verified many of the conceptual best practices that were introduced as part of the literature review phase.

The third and final phase consisted of a survey sent to various executives from other financial services companies. The survey provided feedback about their perception of what data governance consisted of, its significance within the organization and what steps and scope if any they were taking to implement a data governance program.

This researcher’s use of a triangulation of qualitative research phases covered the theory, practical and consensus views towards the subject matter in order to prove that data governance is necessary and feasible to implement within the financial services industry.

1.4 Success Criteria

The three phases of research focused on the definition, the makeup, necessity and feasibility of a data governance program. In meeting these objectives, this researcher focused on delivering the following success criteria.

- Identify the different components of data governance, their purpose and how each contributes to the overall strategy of data management strategy.
- Justify the need for a data governance program within the financial services industry.
- Quantify many of the aspects of data governance, and put a dollar and/or effort figure in place where possible.
- Identify the complexities between the theory and practice of data governance.
1.5 Summary

The evolution of information has been profound in recent years both in utilization and sheer volume. The governance and best practices around information management has invariably changed to cater for the ever increasing demand for more and better data. Chapter two provides a review of literature identifying the core components, practices and obstacles organizations may face when pursuing the data governance program.
Chapter 2 Literature Review

As shown in Chapter 1, financial organizations acknowledge the need for data governance (DG) to reduce costs, mitigate risks and provide regulatory transparency. But the challenge has been the wherewithal to initiate, implement and most importantly maintain governance. Many organizations have initiated projects to clean up “dirty data,” yet without the overarching framework of standards, policy and data lifecycle management, their impact has been low and fails to resolve the root cause of subpar data quality.

The purpose of this chapter is to look at the research that has been done on data governance. More specifically, it explores the impact of low quality data and its underlying costs as a validation for the need (and urgency) of data governance. In defining the meaning of data governance and its best practices, an examination of the key concepts of DG, namely the data governance maturity model, data quality, integration, metadata and master data management (MDM) was conducted. The principles and best practices provided evidence to support the hypothesis proposed in Chapter 1. In addition, the information provided a way to gauge the effectiveness of a data governance program implemented as part of the case study detailed in Chapter 4.

2.1 The High Cost of Bad Data

The impact of bad data comes with both tangible and intangible consequences, the most notable being cost. Most companies in the financial services arena either do not track the dollar cost of data quality issues at an enterprise level or in the cases where the capability exists they do not publicize it. A 2002 Data Warehouse Institute report put a figure of approx USD$ 610 billion on postage, print and rework alone. Redman (2008) pointed out a number of case studies that
estimate accuracy field error rates of up to 5%. Other studies have produced benchmarks of
costs as high as 8-12% of revenue. English (1999) put the cost figure closer to 25% of revenue as
closer to the truth, and claimed that poor data quality is baked into the “normal cost of doing
business” and as such is transparent at an organizational level. However, these numbers reflect
only one of the dimensions of data quality - accuracy and do not consider other areas such as
relevancy, currency and completeness.

The monetary costs and efforts associated with identifying, mitigating and reworking bad
data may pale in comparison to the intangibles such as loss of customer patronage, regulatory
reviews and demoralized as well as high turnover of staff. The fluid nature of data allows it to
permeate through different departments and hierarchies of an enterprise, propagating revenue
losses, forcing higher costs and possible lost opportunities as it transforms and transfers within
the organization. Data quality mistakes abound from Barclays Bank being fined UK £2.45
million (USD $4.07 million) for failing to provide accurate transactions to the UK regulatory
authority (TradeNews, 2009) to a doctor in Tampa who amputated the wrong leg after consulting
the operating room blackboard, schedule and hospital computer system (New York Times,
1995).

Although monetary costs of poor data have not historically been easy to quantify, some
progress in measuring data quality costs have emerged. Redman’s (2008) Cost of Poor Data
Quality (COPDQ) chart as shown in Figure 2.1, categorized the multiple and varied impacts poor
quality data can have from strategic, tactical and operational perspectives. Although not
assigning monetary value to the impacts, COPDQ was effective in creating an itemized list of
workflows that each department/company can assign an arbitrary cost and showing current
operational and downstream strategic impacts.
As shown in Figure 2.2, Ross and Perry (1999) developed another cost approach called the 1-10-100 rule which is used widely in measuring manufacturing defects. The rule is based on the premise that it costs less to prevent a data issue than to retroactively fix it after the fact. The longer the data defects exist and penetrate the organization, the cost and effort to resolve grows exponentially. In a working example, Lager (2009) studied the validity of the 1-10-100 rule and found that companies in the customer relationship management (CRM) realm that follow data governance best practices have a 66% increase in revenue over non-compliant companies. Lager further stated a best in class strategy must be cross-divisional and must qualify data from inquiry to closing the sale.
2.1.1 Sources of Bad Data

The high cost of identifying and fixing bad data aligned with data’s elevation to being an enterprise asset should be enough to eradicate the sources of bad and untrustworthy information. In modern organizations, the opposite is likely to be the case. English (1999) explained that data quality is not a “sexy topic” when compared to closing a sale or executing a trade. The second reason is the acceptance of business of a lower grade of data and the adoption of the “I’ll fix it myself in Microsoft Excel or Access” approach to getting things done. Redman (2008) identified a number of different areas where data quality and governance are called into question:

- **Silo Creation**: Departmental approaches to data integration often result in “having our own copy” of the data. Loshin (2009) referred to them as Line-of-Business silos. Dyché (2006) termed them “creating the new legacy systems.” This lack of enterprise view of data leads to additional costs for application licensing, storage, maintenance and inevitably integration into additional data stores. English (2009) saw enterprise data being an accumulation of interdependent components not to be perceived as
independent units. Silos may also have risks associated with currency where strategic decisions are made based on stale or incomplete data.

- **Duplication:** As data is loaded into various repositories, the data is effectively cloned in either its original or transformed state, increasing the number of versions and locations of the content. In many cases, no lineage between the content outside of the loading process exists. English (1999) emphasized this point in stating that duplication may not necessarily point to identical values, but the duplicate representations of a single real world entity. This loose coupling of the data can lead to questions such as which copy is the authoritative source as well as data life cycle, retention and e-discovery repercussions.

- **Lack of Business Meaning:** Data mis-interpretation is prevalent in most financial services companies. As data moves between departments, information may be misconstrued as each business group applies its own domain interpretation. Dorfmann (2010) provided an example of where common data is leveraged to address risk and financial queries when he wrote, “the problem with P&L and risk data is that it tries to combine two very different views: a mark to market (economic) and an International Financial Reporting Standards view. The future lies in a market data and securities master that enables valuation and reconciliation across the bank.”

English (1999) referred to this understanding as the contextual clarity that intuitiveness of the information and its underlying metadata bring to the understanding of the knowledge worker. In many cases, this information is not well documented resulting in a loss of core knowledge when a subject matter expert leaves the organization.
• **Exponential Growth of Data:** Companies are progressively acquiring more and more data. The exponential growth in data, however, is not counteracted by the disposition of retired or redundant data. This can be attributed to a number of factors such as a lack of retention policy, lack of knowledge on the data domain and a fear of destroying items even if they are passed their end of use by date (Redman, 2008).

New forms of data are also adding to the complexity and explosion in growth, Adrian (2009) pointed to non textual content now residing in data bases such as images, video, music and documents. The growth in unstructured and semi structured content such email is predicted to grow at an annualized rate of 41% (Adrian, 2009).

### 2.2 Data Governance Best Practices

The proliferation of data (or pollution in the case of bad data) throughout the enterprise is something that most organizations need to be address. To achieve this goal of clean, useful provisioning of data, it must be recognized as an enterprise asset with a robust data governance plan implemented. Adrian (2009) recommended a number of steps to manage the growth and pervasiveness of data. They included silo awareness, measured data value and monitored usage, quality and lifecycle management. Achieving these goals calls for a disruptive overhaul of data management and involves a deep analysis of what data a company possesses, how it consumes the information and its information policies. Sarsfield (2009) also questioned the ability to handle change and queries around data ownership and accountability.

Although data may be classified as an enterprise asset, the politics that surround it are unmistakably local. Dyche et al. (2006) raised the question and contention around data ownership being one of the biggest challenges to master data management (MDM). Redman
(2008) listed ownership and brutal politics as No. 1 in his “12 barriers to effective management of data and information assets,” citing their perceived notion of holding power and control. Data sharing is not a natural act as evidenced by asking a stock broker to divulge his stock selection strategy or a sales person to supply their best contacts. Yet, it is only through the promotion of data to the corporate level and out of the departmental silos that one can see its true value (Fisher, 2009). To achieve this objective, a number of data governance best practices should be evaluated and implemented.

The following sections review the prominent components that are critical to any data governance initiative, namely MDM, data integration, data quality and metadata management as shown in Figure 2.3.

![Figure 2.3 Data governance layers (Loshin, 2009)](image)
2.2.1 Data Governance Maturity Model

Many organizations though keenly aware of the need for a data governance program either don’t know where to begin, or see it strictly as a strategic initiative with little or no operational input or value. Smith (2007) acknowledged this strategic long term view, but also saw a number of short-term tactical advantages to DG such as reduced support times or the faster delivery of services. However, the long term objective of an organization should be to build a capability around their information assets to gain trust, protection and leverage this important asset for their competitive advantage.

To attain this level of capability, many organizations turn to a data governance maturity model (DGMM) to provide measurements of current state and projections for interim and future state objectives (Fisher, 2009). The DGMM is based on the popular Software Engineering Institute (SEI) capability maturity model integration (CMMI). The model provides an incremental approach that allows for different assessments of states from unaware through to governed competency. Different researchers and practitioners have identified different states or approaches. IBM (2007) identified six steps in their maturity model, while Gartner (Bitterer, 2007) selected five (Figure 2.4). But all appear to have elements that Fisher (2009) identified in his four groups:

- **Undisciplined**: Fisher estimated that up to 35% of organizations fall into this category. There may be an awareness of data quality issues, but little or no impetus from top level management down to address the issue. Data management in these entities usually consisted of fire fighting, data rework with little regard for the impact of domain level decisions on downstream systems.
• **Reactive**: This group constitutes 40-50% of companies. In these organizations, there is an awareness of the importance of data quality, but more at a domain or departmental level. Data quality technologies may exist, but their use is disbursed with no executive sponsorship or cohesive enterprise strategy for their use.

• **Proactive**: Fisher placed 10% of all organizations into this group. For proactive organizations, there is a realization about the value of high quality data and its governance. Standards have been drafted, but their implementation and adherence may not be fully implemented.

• **Governed (including Managed)**: For organizations that have achieved this level, a full data governance program has been implemented enterprise wide. The program is resourced with the appropriate stewardship, budget and executive level sponsorship. Data is trustworthy and both operational and strategic decisions can be made with confidence.

![Figure 2.4 Data governance maturity model (Bitterer, 2007)](image-url)
The DGMM can prove an effective high level indicator of an organization’s progression towards data governance. The model itself can be further decomposed into the major elements that constitute data governance. IBM (2007) identified 11 elements through their data governance committee as representative of their data governance framework. Key elements included supporting disciplines such as data architecture and metadata, while core and enabling disciplines included data quality, ownership, and lifecycle and policy management. The elements are collected into higher level envelopes representing supporting and core disciplines, enablers and outcomes as shown in Figure 2.5. A DGMM roadmap should include the categories of data governance. IBM (2007) advised that an organization sets a measurement (and assessment) of achievability within each category for progression to the next phase of the DGMM.

The remainder of this chapter examines the core and supporting areas of an effective data governance plan as defined in Figure 2.3, to include data architecture, integration, master and metadata management and data quality. Enablers such as stewardship, ownership, information policy and organizational representation of data are discussed within the context of the core and supporting components.
2.3 Master Data Management

The organic growth of data silos within the organization has led to a number of data quality issues including redundancy, poor integration and lack of entity identification and understanding. As the business evolves, so does each department’s interpretation of the data they use, to the point where there is no enterprise wide knowledge of what constitutes a customer, product, supplier or employee. Master data management (MDM) has now emerged as a technology that can provide single unified views of an organization’s core business entities (Fisher, 2007).

2.3.1 MDM Definition

MDM has become a popular topic in marketing and vendor circles as each group looks at the revenue potential of implementing a new technological framework. MDM however, cannot be solved strictly using technology (Loshin, 2009). MDM and its underlying data management
framework are more complex than a simple physical representation of rows and columns. Operational and strategic decisions must also be considered when making choices that affect an organization’s bottom line in addition to its social and political responsibilities (Redman, 2008). Governance must therefore an integral part of MDM; it must involve extensive collaboration efforts between technical and business resources. Consideration must also be given for external data providers and customers when designing and implementing an enterprise wide authority such as MDM (Sarsfield, 2009).

Dyche (2006) provided a good definition of MDM as a “set of disciplines and methods to ensure the currency, meaning, and quality of a company’s reference data within and across various data subject areas.” She decomposed MDM into four distinct data types:

- **Transaction data:** Here the data is the core operational data that represents an activity or an event at a point in time. Stock trades, net asset value (NAV) calculation on a fund are examples of financial transactional data.

- **Reference data:** Here the data is information that provides the integrity of the business object by uniquely identifying the business entity and differentiating it from other business entities. For example, a future derivative and a money market account are both investment vehicles, but are distinctly different assets.

- **Relationship data:** Relationship data shows the business representation of the data and how it constitutes the business entity (structural relationship) and relationship to other business entities (object relationship). In the financial services industry these structural compositions and relationships are core differentiators, determining what instruments can be traded, their makeup and the
relationships between the traded entity and other financial entities such as issuers and counterparties and financial remuneration. It is this full 360 degree view of the entity, its content, structure and inter-relationships that provide the business value of information.

- **Metadata:** Metadata is “data about data,” it provides descriptive markup information about data elements. The term, however, is a little misleading according to Loshin (2009) as it does not give proper recognition to the importance of metadata to the whole data governance equation. Khatri et al. (2010) identified different types of metadata such as technical, business and domain specific metadata. Both have described metadata as the glue that connects a database to the real “world”. It can define technical structures, data representation, and lineage and information policy. Redman (2008) saw metadata as critical and defines it as among the most important and often overlooked data components in the organization.

Many terms have been used to describe MDM. These have included “single version of the truth,” reference data management, authoritative source and “golden record.” Each definition, however, takes a simplistic view of what MDM does and does not provide. Single versions of the truth are challenging. Examples include different departments having different perspectives of business entities, a representation of a customer to the salesperson may not be an exact match when compared the legal and helpdesk department’s perception of their clients. A technological perspective on the single version may also differ significantly. Despite the organizational and political challenges to MDM, Andreescu and Mirce (2008) pointed to certain
key drivers such as the improvement in transactional management, cross enterprise view of business data and the ability to provide better data governance. The next section will explore these drivers in more detail.

### 2.3.2 Benefits of Master Data Management

Master Data Management is not a standalone project, but an ongoing program with iterative steps to ensure that data is constantly analyzed for quality, matched, merged and published. The effort to justify, initiate and maintain MDM data is significant, yet the payoff can also be extensive. Loshin (2009) highlighted the following seven benefits:

- **Increased business information knowledge**: Having full 360 degree knowledge of one’s business domain allows for better operational and strategic decision making. English (1999) pointed out that the knowledge worker needs access to modeled data, shared learning and continuous improvement to thrive in their positions.

- **Meeting regulatory compliance**: The economic meltdown of 2008 has led to additional regulatory oversight of the financial services industry. The Dodd-Frank Act, Volcker rule and existing regulations such as Sarbanes-Oxley, Basel II and payment card industry (PCI) compliance have made transparency within the financial services industry an imperative. MDM helps provide a unified view of any regulated business entities, their data composition and any policy related and applied across systems. Bhatt et al. (2010) and Dyche (2006) placed emphasis on the fact that regulation can be global, federal or local in nature, sometimes with conflicting requirements and in many cases duplicated action items. They highlighted a number of challenges when it comes to regulation such
as conflicting business definitions, missing or inaccurate data and the lack of metadata and ownership. Loshin’s (2009) addressed all these issues within the governance layer of MDM.

- **Enhanced business productivity:** Two of the biggest impacts on productivity are error handling and root cause analysis. Ammerman (1998) defined root cause as a process to systematically detect and analyze problems to determine and apply corrective action. Using the 1-10-100 rule discussed earlier, a single defect can consume multiple resources in investigative and remediation tasks for considerable periods thus reducing their productivity within their primary function. Continual firefighting can demoralize resources. English (1999) identified this as a key contributor leading to disinterest and employee turnover. Productivity also suffers when knowledge workers cannot easily find the appropriate business information. Redman (2008) claimed that most knowledge workers spend up to 35% of their time looking for information to perform their business function and in over 50% of these cases they fail. MDM’s standardized and unified view of business entities aligned with a robust data governance layer provides the necessary views, findability techniques and policies to alleviate much of this effort.

- **Major competitive advantage:** No two company’s data are exactly the same. How and where an organization uses this information can provide a distinct competitive advantage over its competitors. Carr’s famous 2003 article “IT Doesn’t Matter” points to the commoditization of hardware and software. The advent of cloud computing, outsourcing and open source applications confirm Carr’s prophecy, eroding any competitive advantage gained that may be gained from superior infrastructure. Carr stated that only proprietary and intellectual properties that can add to valuation and to net worth should
be supported. Redman (2008) agreed and noted that corporate data meets all these criteria and that it uniquely identifies anything and everything that matters to an organization. Business intelligence and data warehouse technologies may be deemed strategic and adding to competitive advantage, but without well defined, authoritative and high quality data populated, they are merely another silo of data to manage (Dyche, 2006).

- **Improved data quality**: MDM and high data quality are tightly coupled. MDM’s success depends upon having extensive data quality processes to ensure the integrity of data for downstream systems. Without quality, Loshin (2007) saw MDM being just another “island of information” defunct of any capability to provide competitive advantage or regulatory controls. A centralized base for applying data quality rules also leads to better controls and reduction in development and maintenance costs (Dreibelbis, 2008).

- **Reduced costs**: A Gartner 2008 research survey (McGriff, 2010) predicted that by 2012, MDM will have reduced or eliminated redundant master data by 60%. Redundant or low quality data costs money, the cost of inefficient data can be crippling from a business perspective in both dollar amounts and lost goodwill (Redman, 2008). The technology overhead is also substantial. Unfortunately, the perception of disk storage being cheap is a misnomer, when one considers the other overheads such as maintenance, query performance, retention policy and the risk of regulatory e-discovery and its implications (Loshin, 2007). McGriff (2010) identified MDM’s consistency and top down view of risk as one of the biggest benefits towards risk avoidance and costs.
• **Improved customer service:** In the highly competitive world of financial services, customer attrition rates can be higher than other industries. Prices and fees have become standardized for the most part within the industry, making organizations search for other features and services to maintain competitive advantage. Dyche (2006) pointed to fully knowing and understanding the customer, their behaviors, trends and dislikes as a key differentiator. In a modern organization, most of the customer information resides in multiple systems where obtaining a unified view of the customer would be impossible without an MDM solution.

MDM and CDI systems bring a lot of advantages to the organization, but they are not without their faults. Ambler (2003) described the unified view of “one size fits all” as not being realistic from a multi domain business perspective. She provided some additional points:

• **Program abandonment:** The constant review and management process involved in MDM and governance can become overbearing. Over time, this can lead to both the technology and business resources being discontent and move to the possible abandonment of duties. Redman (2008) advocated a carrot and stick approach with the use of monitoring to ensure compliance and an incentive structure (possibly built into a performance review) to reward compliance.

• **Politics:** Who owns the data? On the one hand, business people point to the technology’s application and storage and designate the information technology (IT) department. On the other hand, IT refers to the business semantics, regulatory and the operational and strategic use of the data to route ownership towards business. Fisher (2009) believes that both parties are responsible for the welfare of the data and its management at an
enterprise level. Within an MDM system, there is plenty of opportunity for politics with accountability for many facets of the system. A modern MDM system includes business and technical ownership, data stewards and knowledge workers across many artifacts such as business terms, processes, applications and business entities (Tozer, 1999). MDM is a complex topic and will be addressed in more detail in the MDM Ownership section.

- **Moving data needs:** The repurposing of data to meet additional requirements may call for a whole restructuring of the data model. Ambler (2003) believed achieving the “one truth” is a futile exercise and that organizations should adapt a “similar truths” approach. The adoption of data federation over unified views is also encouraged. Federation allows each domain some flexibility to build their “similar truth” representation without being forced to use an inflexible model or resort to data silos.

Master data management offers some compelling benefits but also comes with its complexities and risks. The next section will look at some of the different architectural techniques and structures used to accomplish MDM.

### 2.3.3 MDM Structure

Achieving a single version of the truth provides a number of architectural challenges. Related business data may be dispersed among many systems such as customer relationship management (CRM), enterprise resource planning (ERP) and legacy applications. To provide a unified view of data, the MDM architect must decide whether to query the core data elements that constitute the business element from multiple systems or a more persistent design is in order.
Gartner (Radcliffe, 2006) defined a number of different MDM hub patterns to meet the varying needs of different organizations.

### 2.3.4 Registry Hub Patterns

MDM Registry Hubs provide a virtualized view of the master data. In the registry structure, only identifying data attributes are stored and a metadata data retrieval layer is used to collect and federate the data (Berson et al., 2007). Dreibelbis (2008) acknowledged the benefits of retaining identifying only data but also makes a case for keeping additional attributes for consistency, relationship and identifier quality checks. Registry hubs provide real time access to data while eliminating data migration and redundancy. The information can be transformed “en route” to meet the master data representations. Quality checks are limited to the integrity checks around the data identifiers leaving the extensive quality checks to the source systems Dreibelbis (2008). Berson et al. (2007) suggested this model in environments where robust and established source systems exist or where real time access is an imperative.

Dreibelbis (2008) pointed to its low cost of entry as a plus and saw its use for transactional based operations over more analytical support systems. Dyche (2006) referred to the registry hub as the “reference style” hub and sees the biggest benefit being it's “on request” approach to data acquisition. Registry hubs do not come without their faults. Dyche (2006) identified performance as the main culprit and the lack of caching with full replication needed for duplicate requests. Berson et al. (2007) refer to the limitation of read only access and also the lack of a data quality and survivorship rules as further constraints. Dreibelbis (2008) saw the tight coupling of multiple source systems as an issue, citing the real time dependency of the MDM on all source systems as a major vulnerability as shown in Figure 2.6.
2.3.5 Coexistence Hub Patterns

In certain situations, having real time access to master data is not the most important factor. A marketing campaign selecting high net worth investors would consider yesterday’s data current and fit for purpose. In these cases, data can be passed from the source system in an acceptable window to meet referential master data needs. The term given to this design is coexistence hub patterns, as the data exists in the source system and also the MDM hub (see Figure 2.7). In coexistence design, the source system still maintains the responsibility of being the system of record, while the MDM remains the master data point of reference. Dreibelbis (2008) identified coexistent hubs as having convergent consistency, as the currency of the MDM object is dependent on the synchronization windows of the constituent parts.

Researchers agree that many benefits exist when using a coexistent hub. Dyche (2006) cited the persistent nature of the data within the hub and to the fact that the data is localized and transformed to the correct MDM representation. Loshin (2009) saw it as a happy compromise between the “thin” registry hub and the “thick” transaction hub. He also referred to the master
Dreibelbis (2008) noted the improved performance of having persistent data and the ability to centralize data quality checks is also significant. Dyche (2006) described the added complexity and the synchronization schedule to get a most current view of the MDM object. Additional storage costs are also required for this architecture as data is replicated to the hub.

Dreibelbis (2008) cited the fact that hub data can be updated and or deleted within its domain context, but that any further updates on the source record could lead to data conflicts.

**Figure 2.7 MDM coexistence hub (Dyche, 2006)**

### 2.3.6 Transactional Hub Patterns

Loshin (2009) put transaction hubs into simple terms, citing that it provided a single repository to manage all aspects of master data management. The MDM system becomes the authoritative source and also the reference system of record for its domain business entities (Figure 2.8). In a transaction based hubs, read and write operations can be done within the MDM
environment or by source systems using a certified web service (Dreibelbis, 2008). Absolute consistency is also attained as there is no synchronization dependency related to the master data. Transaction hubs represent the materialization of the master data within the MDM, therefore the trustworthiness of the data needs to be established through rigorous quality checks being placed on the golden record (Dreibelbis, 2008). Transaction hubs are the correct choice when organizations do not want to invest additional budget in source systems where there is a definitive and logical benefit in managing the system record centrally in an MDM hub (Berson et al., 2007).

In the case of a transaction hub, the dependencies are reversed with the source systems reliant on authoritative data from the hub. Loshin (2009) and Dreibelbis (2008) both defined this design as the most disruptive and costly to implement. Data source systems and archaic legacy systems need to be altered to facilitate consuming published MDM data. Dyche (2006) noted that most enterprise resource planning, customer relationship management and business intelligence vendors propose transactional MDM as a best practice especially when the enterprise owns existing products from the vendor.

![Figure 2.8 Transactional hub (Dyche, 2006)](image-url)
2.3.7 External Reference Data

In addition to internal data sources, the financial services industry is a huge consumer of external data; sourcing rating, risks, performance, trade and customer data from many trusted third party providers. Strategic information providers to the financial services industry include Bloomberg, Factset, Morningstar and Dun and Bradstreet. Gartner (Radcliffe, 2006) identified externally provisioned data as critical to the organization and should be verified where possible to an external reference database. Many of the characteristics of the external reference mimic that of MDM (i.e., to provide to uniquely identify a resource and have complete and upmost trust in the creditability of the data and the source).

2.3.8 MDM Selection

MDM architecture selection is not an easy decision. After gaining executive buy-in, the MDM team has to analyze a number of different criteria to come up with the best approach. Loshin (2009) identified some items to consider being system rationalization, synchronization, security access, maintenance, maintenance effort, MDM objects and attributes collection and inevitably cost to implement and maintain. He provided a complexity scale showing the gradual escalation of effort based on level of consolidation and attribute collection. Figure 2.9 shows these findings and positions the different flavors of MDM in terms of their complexity and coupling to the source systems.
Dreibelbis (2008) and Berson et al. (2007) pointed to a “phased approach” and natural progression from registry hubs through coexistent and onwards towards full fledged transactional MDM.

Although data is viewed as a corporate asset, researchers advocated that there must be ownership of the information at a more granular level within the organization. Dyche (2006) believed it is the most prevalent question when implementing data governance programs, “Who in the company should own the data?” Berson et al. (2007) did not draw departmental lines, but established the owner as the resource or group that is in the position to make strategic decisions about the data, including quality, security and access considerations. Redman (2008) viewed data ownership a dangerous notion when depicting power and rights that can be used for both progressive and resistive motives (with the second coming more naturally). He noted that if data ownership is poorly defined, brutal political storms and debates about personal and organizational power and control can result. Dyche (2006) reasoned that the lack of ownership and resistance to share data might be the aversion to having to subsequently inherit the

![Figure 2.9 MDM hub complexity (Loshin, 2009)](image-url)
accountability and responsibilities for the data’s quality and issue resolution. Many see this as a hindrance and not adding any additional value to their domain of knowledge or wealth.

As researchers have noted the question of data ownership is a complex one which is compounded by additional responsibilities such as regulatory control and reporting. Identifying data owner candidates using a set of criteria was promoted by Loshin (2002). Data within the organization can have a number of touch points such as creation, consumer, licensee, packager and compiler to name a few. Domain and geographical considerations also need to be acknowledged. For example, an institutional investment account may differ structurally, regulatory and in representation based on its geographical location. Loshin (2002) recommended getting a full picture of the “paradigm of ownership” to ensure the correct person or group are assigned and are comfortable with the responsibilities.

2.3.9 Roles and Responsibilities

“The data will not govern itself.”


To help minimize conflict over data ownership, MDM has broken down the ownership title into business and technical owners, data stewardship and subject matter experts (SME). This scheme acknowledges that MDM is a multifaceted discipline where the data owner may not be the person responsible for maintaining and interacting with the data on a day to day basis. Successful MDM must have assigned resources and each must understand and commit to running the initiative as an ongoing continuous program. Failure to maintain all the components of the MDM framework can lead to creditability issues and to its ultimate demise. Each group within the enterprise has a vested interest in ensuring that they are receiving the best data
Achieving this goal requires time, money and resources in the form of designated roles and responsibilities.

One of the most important resources to identify early is the data steward. The data steward or knowledge worker (Redman, 2008) should offer subject matter expertise around the data and its use. Seiner (2006) argued that stewardship has accountability over data and as such recommended a formalization of the role, rather than using an implied assignment. In an earlier study, Seiner (2002) believed that these resources already exist in the organization and that it’s a matter of using the 3-D’s (De-facto, discipline and document) approach to identify and formalize their role. Karel (2007) saw the role as so critical to MDM that the individual or group’s responsibilities should be part of their yearend performance appraisal. Karel also stated that during initial rollout that 1% of project budget should be allocated to initiating a stewardship program.

Data stewards, though important, are only one component of the MDM / data governance equation. Loshin (2009) identified additional resources including executive sponsorship, data councils to mitigate risk, business consumers and technical support staff to maintain the program. Loshin (2009) and McGilvray (2008) argued for using a RACI chart as shown in Table 2.1 to identify the different groups and roles and to detail the degree of accountability the different resources have for each data management event.
Dyche (2006) and Fisher (2009) advocated the early establishment of a data governance council, as a critical part of overall governance. The board of directors should be a cross representational group of business and technical resources with the power and conviction to set enterprise wide data standards, definitions, acceptable quality and policy for information across all lines of business. Dyche also recommended that the data governance council should have teeth by reporting upwards to the executive branch and managing downwards to ensure governance is maintained. Sarsfield (2009) believed adding a third party advisory was also beneficial as it gives an outside independent voice. The external consultant may also act as a mediator or voice of reason in contentious situations.

2.3.10 MDM Summary

Master data management and its underlying data governance vehicle can offer many benefits to financial services companies. Most organizations realize the benefit, but fail to grasp the planning and management effort and associated costs. More importantly, the false assumption that MDM is a technology initiative must be overcome. MDM is not a project, but an

<table>
<thead>
<tr>
<th>Task</th>
<th>MDM Management</th>
<th>Business Clients</th>
<th>Application Owners</th>
<th>Information Architects</th>
<th>Data Governance</th>
<th>Metadata Analysts</th>
<th>System Developers</th>
<th>Operations Staff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Develop data harmonization processes</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>Data requirements analysis</td>
<td>I</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>R</td>
<td>R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metadata analysis</td>
<td>I</td>
<td>C</td>
<td>C</td>
<td>R</td>
<td>R</td>
<td>A</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Master data models</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>A</td>
<td>C</td>
<td>R</td>
<td>C</td>
<td>I</td>
</tr>
</tbody>
</table>
enterprise wide program that must have the human capital investment assured to start and maintain the effort.

### 2.4 Data Integration

The movement of data throughout an organization is a pivotal component to success. Both operational and strategic initiatives rely on having the correct data at the right time and in the right place to make decisions. Redman (2008) went a step further by claiming that it is only through movement that data is actually creating value.

#### 2.4.1 Why Do We Integrate Data?

Data at rest creates little or no value. Data has value when it moves from location to location for a business purpose (Redman, 2008). The business reasons for integrating data vary. Examples include data warehousing, business data repurposing, new software implementations and upgrades or business process automation. Financial services companies move vast amounts of data on a daily basis. In many cases, this information is either sourced externally from a content provider such as Bloomberg or Factset, or internally from departmental or other diverse sources. A typical operational data flow may see trade confirmation data being leveraged by portfolio management, fund accounting, sales and marketing as well as risk and compliance. Redman (2008) stated that data does not always follow a predefined path and that it is constantly moving through an organization. Each department may have its own interpretation, business practices, rules and demands of “their” information. In many cases these may not be consistent with the overall corporate policy.

Data integration is the term commonly used to represent the movement of data; the phrase represents a myriad of technologies and methodologies from legacy COBOL applications,
batch, real-time and the currently popular web-services. In this review data integration as whole, rather than individual technologies are examined, characteristics, functionalities, risks and governance concerns around data integration are highlighted.

2.4.2 Defining Data Integration

"Don’t pass bad data on to the next person. And don’t accept bad data from the previous person."

Redman (2008)

Data integration defines the processes, standards and controls needed to source, transport, transform and insert data to a target repository to fulfill a business need. Although many different technologies can be used to apply data integration, the stages of an integration process are standard. Figure 2.10 depicts the four different stages of typical data integration:

- **Extraction**: Data extraction is the initial step in any data movement. The retrieval process is based on a business requirement to have information made available in a target location for use. The source and target locations can be database stores but may also be files, reports or web services. The objective with extraction is select only the information that is required, it should represent of a real world business entity it must maintain consistent meaning between source and target systems either directly or through transformation.

- **Transportation**: For decades, paper data files were the primary vehicle of data transportation. Data was extracted from source systems into a formatted text file and was loaded into target systems. The introduction of databases, enterprise
service bus and web streams have provided greater transportation flexibility. But it also has added to the complexity and maintenance overhead for data movement.

- **Transformation:** In many cases, data from the source system is not a direct match for that at the target location. The data integration process therefore takes responsibility for implementing the value mapping, transformation and integrity checks needed to load the data set. In many cases, the consistency and meaning between the source and target systems is lost in translation. Loshin (2009) called for a service layer within master data management that provides the conceptual master data objects, meanings, representational values and standards (business rules) around data usage. Dyche (2006) and Maydanchik (2007) in a similar stance advocated the use of logical and conceptual models as an important differentiator from the physical model upon which most many data integrations are built.

- **Insertion:** The insertion (aka load) process is the final step in the data integration process and represents the insertion of data into a database, a report or a web service. All data quality, business rules and target representations of the data have been completed and the information is ready for consumption.

![Figure 2.10  ETL process and data (Radcliffe, 2009)](image-url)
Integration as a process has a finite number of steps, yet it represents one of the most complex and higher risk areas of data management. Governance should be at the forefront of any data movement as any integration has the capability to introduce quality degradation, redundancy and possibly other data governance risks. Dyche (2006) raised this risk in stating that the number of data movements is inversely proportional to the quality of the data. Cali et al. (2004) highlighted a number of integration issues and their impacts including data quality inconsistency between the semantic interpretation of data at source and integrated downstream systems. A DataFlux (2010) white paper identified many of the challenges facing organizations and their attempts to integrate disparate data repositories as shown in Figure 2.11, the top two being data quality and funding. The tangible value of data to an organization would have to be questioned, when doubtful quality, underfunding and lack of security around one of its most prized assets is considered an acceptable practice.

![Top Data Integration Challenges](image)

**Figure 2.11** Top data integration challenges (Dataflux, 2010)
Many of the risks associated with data integration can be correlated to the data quality dimensions addressed in the data quality review section. There are however some additional process risks identified.

- **Business Representation:** Redman (2008) pointed to data inconsistency and poor data definition as two of his seven common data quality issues. Poor data definition can lead to operational confusion. Loshin (2009), Dyche (2006) and Smalltree (2006) each pointed to the need for logical data models at an enterprise level as a critical component of any MDM initiative. In too many cases, the logical model is neglected in the race to create the physical model and start development. Levy et al. (2004) promoted the use of data profiling of data as a continual improvement practice and cited the fact that not only data changes over time but so does the higher level informational needs of the business entity.

- **Development Framework:** Dyche et al. (2006) differentiated integration for MDM from data warehouse and other analytic and operational stores. They cited the function oriented approach for MDM required it to meet strategic (top-down) and operational (bottom-up) approaches to integration development rather than a point to point solution.

- **Authoritative Source:** Brewer (2006) identified having a trustworthy source for business elements as one of the critical elements of identification management for Sarbanes-Oxley (Section 404 IT Compliance). In addition to regulatory policy, a data integration best practice requires that all target systems should source their data from the trusted and certified source. English (1999) titled the authoritative source as the “Record of Reference” upon which he believed considerable data
quality must be applied. English also called for an accuracy to surrogate source rule and an important check to ensure that the data in the surrogate system remains consistent with the trusted source. Redman (2008) pointed to the availability and knowledge of trusted sources to the knowledge worker. Authoritative sources are ineffective if they are not known, updated and published. Resources need to know which system provides the trusted data for customers, products and employees. The knowledge and availability of trusted data should therefore be leveraged whenever possible for data propagation to downstream systems.

- **Uncertainty Management:** Magnani and Montesi (2010) highlighted the fact that integration between disparate sources will always bring uncertainty. The causes can range from ambiguous schemata, ill calculated aggregations to uncertain mappings between equivalent entities. Data integrations have “extract knowledge” of how to map source data to integration schemas, Dong et al. (2009) argued that as the integration scope broadens both complexity and uncertainty are introduced. In some cases, uncertainty is inevitable such as when a trade or order has been executed. The event cannot be rolled back to clean the data. In such cases Magnani et al. (2010) recommended an uncertainty management framework using a combination of fuzzy logic, evidence and probability to identify dubious transactions and rectify the outcomes post execution.

- **Performance:** Having the right data at the right time to make the right decision is an imperative in today’s business. Berson et al. (2007) discussed this as an important dimension for organizations as they move from account centric
approaches to a more unified customer centric view of business data.

Traditionally, data integration in batch was the norm, but in world of high velocity trading, real time quotes and risk exposure there is a requirement for a more rapid delivery of information. Loshin (2009) broke down performance based on different “needs”, the analytical versus transactional need, acceptable latency, coherence needs (predictable synchronization) between systems. The synchronization of data is only as immediate as its weakest link, today’s integrated environments leverage many different architectures in sourcing data such as relational databases, web services and venerable COBOL files and copybooks. Different architectures may have their specific optimization techniques. Ballard et al. (2009) recommended partitioning, indexing, mixed workload performance features and compression techniques for RDBMS systems. XSLT performance can be increased using caching and template instantiation optimization techniques (Kenji, 2005).

Maintaining a timely delivery of data requires constant monitoring and analysis of the integration processes. Trends such as MDM, volume growth and data repurposing will continue to put more emphasis on information delivery in the future (Dyche, 2006).

### 2.4.3 Data Migrations

Business progression often leads companies to implement or upgrade information systems as means to consolidate their data or gain competitive advantage. Mergers and takeovers also provide opportunities to migrate data during system rationalization initiatives. Regardless of the business driver, the migration of information as an event incorporates a large transition of data from one source to another using many “once off” conversion scripts to implement the
many complex and risky data migration steps. A 2007 SoftTek (IBM, 2007) survey found more than 75% of respondents claimed some to many issues during data migrations including unexpected downtime, data corruption, missing or lost data.

The term “garbage in, garbage out” synonymous with the migration process, provides a baseline for the impact of an ill planned data migration strategy. The conversion of data is a complex process that cannot make any assumptions around the quality, integrity or trustworthiness of the original data. Halpin et al. (2008) suggested that the migration team start with the business domain and identify the main objects which he calls the “universe of discourse.” Halpin further believed only when there is a clear understanding of these (from both the source and target systems) that a migration process should be allowed to proceed. Maydanchik (2007) argued that quality of the data in the target system was directly proportional to the time spent planning and analyzing data. He recommended an 80% effort on analysis and a 20% dedicated to conversion script development. Leon (2008) identified data migration as the single largest task in any enterprise resource planning implementation and advises taking an iterative approach to data migration and using it as an opportunity to weed out unused data.

The 24/7 nature of business systems has narrowed the window for large scale data migrations, while raising the risk of business impact. IBM (2007) proposed a three phase approach to data migration to increase the probability of success.

- **Plan:** The plan calls for knowledge of all component parts including source and target systems, architecture and hardware performance and the initiation of a test plan.
• **Migrate:** Hardware and software certification should be obtained and any script customizations should also be complete. The use of an iterative approach to migration encouraged, with dependency checks as well as pre and post validation checks.

• **Validate:** Although the representation and structure of the data may change, the business purpose of the data should remain the same. Statistics, testing and both technical and business knowledge works should be used to ensure the integrity of the data and more importantly the underlying business function.

From a data migration/integration perspective, Dyche (2006) saw data as “always a bridesmaid” and declared integration as the number one headache of most customer focused initiates. Data is therefore looked at as an afterthought when compared to the more exciting features, functionality and architectures to be examined when implementing new systems.

### 2.4.4 Data Federation

The business representation of data along with Wang’s (1996) timeliness dimension of data has led businesses in a number of different integration directions. Both data integration/migration and the siloed approaches have associated risks, dependencies and costs attached, yet until recently they were the only viable approach. In order to achieve this customer centric view of a business entity as advocated by Benson et al. (2007), data may have to be collected from many disparate systems, each of which may have its own underlying architecture and idiosyncrasies. Web services offer a solution but may be constrained by volume and latency issues for large scale data manipulation. In recent years, the popularity of MDM and customer data integration systems (CDI) has pushed the concept of data federation to the forefront (Dreibelbis, 2008).
The objective of a data federation is to bind two or more heterogeneous data sources without data movement and consequently reduce or eliminate data redundancy (Sauter et al, 2006). As shown in Figure 2.12, the federation pattern is carried out by a virtualization middle layer which handles the requests for data, the decomposition and execution of retrieval statements and the consolidation of results sent to the requesting system.

![Data federation overview](image)

**Figure 2.12 Data federation overview (Sauter, 2006)**

Researchers, however, have noted that data federation is not without its own issues. Sauter et al. (2006) recommended careful consideration around source dependency, and distributed security, latency and performance. Similarly, Kimball et al. (2008) pointed to the overhead of cross-system communication and performance for different concurrency needs of the member data sources. Dreibelbis (2008) cited data quality as an issue stating that the required quality in the source system may not meet the requirements of the federated user. He also acknowledged the query ability of data federation as a benefit but questioned CRUD (Create, Read, Update and Delete) capability to post back to multiple disparate systems.
Loshin (2009) and Dyche et al., (2006) identified data federation as an important part of Enterprise Application/Information Integration (EAI/EII) systems but a critical part of any registry style Master Data Management system where only identifying data attributes are persisted and all other data is sourced from their authoritative source systems.

2.4.5 Data / Information as a Service

The popularity of the internet and the growth of distributed middleware has led to growth of different types of services among them Data as a Services (DaaS). Financial services companies have been receiving data from third party vendors for years, but the difference with DaaS is the dynamic integration of both internal and external data into a logical service that can be consumed by the business.

A 2006 Forrester Research survey (Gilpin, 2006) found what most businesses wanted in their information delivery solution was lower cost solutions, the provision of real-time information, quality, agility, automation, high availability and the removal complexity. The requirements though extensive never state the need for data to reside in a database or be located within the confines of a corporate infrastructure. Kobayashi-Hillary (2007) pointed to a growing global market for sourcing everything from technology, software development and human resources. It is becoming inevitable that the delivery of data can benefit from the same delivery model. Forrester (Yuhanna, 2010) proceeded to define DaaS as a “comprehensive strategy for the delivery of information obtained from information services, following a consistent approach using SOA infrastructure and/or Internet standards. The information delivered may be required to conform to a common information model.”
The Forrester definition is comprehensive and includes strategy, standards and infrastructure to meet these objectives. The following section will discuss the critical components for implementing an Enterprise Information Integration (EII) strategy to deliver DaaS.

One of the primary objectives of DaaS is the removal of complexities in delivering information. Chandras (2005) advocated a federated solution as part of an overall EII plan to break “through traditional barriers of location, structure, semantics and context.” Howard (2005) saw DaaS as a fourth tier where MDM and data federation sits between the data sourcing layer. Data is gathered, transformed and modeled into a logical business services which are then routed the different applications for specific business rules application. Chandras (2005) and Bishop (2008) cautioned that current infrastructure (and architect mentality) was not built for federated service models. Both researchers advocated a paradigm shift in planning and architecture, moving away from extract, transform load (ETL) to a source, model and provision framework. Bishop (2008) concurred and proposed a top down design method that focused on business information demands which are then correlated with “information resources” through service oriented architecture (SOA) based layer. Bishop’s motivation was the sourcing and referencing rather than storing of information for a business purpose, abstracting retrieval, complexity and information location.

The paradigm shifts proposed by Bishop, Chandras and Howard relied on a number of critical components and technologies:

- **Data virtualization and federation**: The virtualization term is used to describe an abstract layer that can be applied to data located across disparate data sources (databases
or web services). The data is gathered, modeled and presented as a single unified logical
data source (Bishop, 2008). Khanna (2008) saw the virtualization of data as invaluable,
reducing departmental silos while also promoting data availability to a myriad of
applications. Loshin (2009) pointed to the fact that most MDM systems do not contain all
the information needed to materialize the “golden record.” In many cases, a registry style
MDM is proposed leveraging a federated model to gather information from disparate
systems.

- **Cloud computing:** The Cloud phrase as been used to describe everything from Software
  as a Service (SaaS such as ADP/Salesforce.com), infrastructure outsourcing (such as
  IBM/HP data centers) to storage and hardware provisioning services such as Amazon’s
  EC2 stack. The true definition of cloud is not important if one is to take Carr’s (2003) IT
  Prophecy into consideration. What is significant from an internal resourcing perspective
  is the ability to gain competitive advantage or conversely identify the best commodity
  vendor to provide the service. Regardless of location Bishop (2008) called for the fast,
  reliable and dynamic access to the most frequently accessed business data.

- **Web services and middleware:** The potential of SOA is well established as an
  interoperable software style for provisioning flexible and reusable services. SOA’s loose
coupling and low cost assembly reduce complexity and cost while using a best of breed
approach for providing services. Middleware also plays an important part in DaaS,
providing distributed services, load balancing and high volume transaction processing
and monitoring. The benefit of using SOA for DaaS cannot be underestimated, yet White
(2005) called the investment a complete waste unless there is the enterprise data to
exploit the capability. For this reason, Dreibelbis (2008) saw MDM as a key enabler of
SOA to reinforce integrity and halt the propagation of low quality and untrusted data to an even larger audience.

- **Metadata**: A key aspect of any integration layer is the markup data known as metadata. Chandras (2005) pointed to its critical use in the retrieval and federation process while he also saw reusability as another area of benefit. Gilphin et al. (2006) used the phrase “You can’t reuse what you can’t find” in describing the risk of needless service duplication. Gilphin saw the extensive use of service metadata as the only way to avoid SOA chaos.

- **Data as a Service**: This is a relatively new concept, but one whose time has come (Yuhanna 2010). Gilpin et al. (2006) coined the term “information fabric” to refer to the business centric virtualized view of information, enabled through an SOA layer and devoid of the technological complexity. Chandras (2005) saw the days of “laissez-faire” data integrations as something that needs to be put in the past, but it is only through EII and an extensive information governance plan that DaaS can live up to its potential and become a reality.

### 2.5 Data Quality

“Right first time” and “Fitness for use” are two of the more popular phrases coined to describe the art of quality assurance. The terms are very relevant in the manufacturing domain, but are they just as applicable to data quality and more specifically data governance within the financial services industry? This section discusses the different properties of data, its morphing capabilities and the complexities of applying data quality standards to this valuable asset.
2.5.1 Data Quality Complexities

As a raw material, data is distinctly different from its manufacturing counterparts. Information is not a “use once” commodity; it can be cloned, moved, transformed and repurposed. Multiple domains and contextual structure of data also adds to data quality complexity. Each internal domain may have its own semantic understanding of the data for its own business use, yet be oblivious to the enterprise meaning, use and quality expectations of the same data. Data can also be sourced from multiple providers, each of which may have their own interpretations of quality, or it may also be distributed externally to meet fiduciary, financial and regulatory requirements.

Although data quality possesses some unique attributes that differentiates it from its manufacturing counterparts, similarities do exist. Wang (1998) leveraged Deming’s quality cycle in his TDQM (Total Data Quality Management) for quality enhancement to identify, measure, assess and improve data quality in a planned, incremental and iterative movement as shown in Figure 2.13. In a similar vain to manufacturing best practice, Redman (2008) pointed out the need for quality around the data supply chain, while both Loshin (2009) and Wang (1998) depicted the collection and organization of data into a product that fills a real world customer need.

![Figure 2.13 TDQM quality cycle (Wang, 1996)](image-url)
2.5.2 Data, Information and Knowledge

Lee and Strong (2003) asserted that knowledge is a quintessential ingredient in any work environment, from a simple data entry to complex derivative trading. Knowledge though is not something that is instantaneous, static or commonplace, but something that is gathered, adapted and assimilated over time. Capurro (in Zins, 2009) defined knowledge as the general understanding gained from the accumulation of information, when allied with experience and context, it can provide breadth and depth and possible new perspectives around existing or new work.

Data and information as terms are often used interchangeably, yet they can have separate meanings. Capurro (in Zins, 2009) defined information as something that is conveyed, that has meaning within its domain that can be interpreted, analyzed or measured. Data conversely he defines as atomic units without context and therefore lacking the characteristics to be classed as information. Childers (in Zins, 1999) refuted the logical hierarchical representation and progression of data through information forming knowledge, yet regardless of structure and bindings there is irrefutable evidence (Loshin, 2001) of the coupling and impact of data and information quality to the collection, validity and continual adaption of knowledge. Redman (2008) went as far as calling data the “means by which organizations encode knowledge.”

2.5.3 Data Quality Dimensions

There are many entry points in an organization for deficient data, from unintentional user error, programming bugs through data misuse and the fraudulent abuse of information. Identification of the root cause, measuring the impact and mitigating the risk are all common steps in most operational workflows. Unfortunately in many financial services organizations, the steps tend to be reactive in nature, leading to operational delays and possible dependency issues.
with downstream systems. A more proactive approach to data management aligns data quality along a number of quality dimensions. Although no definitive creator of the dimensions could be found during research, Wang and Wand’s 1996 paper “Beyond accuracy: What data quality means to data customers” highlighted the need for a broader validation focus on different additional aspects of data over and above narrower focus on data accuracy. Additional attributes such as timeliness, completeness, reliability and relevancy are introduced and explained. Follow-on work from most notably English (1999), Redman (2001), McGillivray (2008) and Loshin (2010) provided additional scope and elaboration to the initial quality attributes, possibly reflecting varied academic and work experience perspectives. Each author used different categorizations for housing data quality dimensions, yet the core attributes persisted in each expert’s methodology. In the following paragraphs, the core dimensions are discussed.

2.5.4 Accuracy

The accuracy dimension is one of the first and most formidable quality attributes defined. Olsen (2003), though crediting the other aspects of data quality, promoted accuracy as the single most important. He described it as “the most visible and dramatic dimension of data quality.” Data accuracy has been defined as how representative data is of “real life” objects. Loshin (2006), Khatri and Brown (2010) referred to recorded data’s conformity to actual value while Olsen (2003) elaborated on this to define accuracy as data values of an object that must be correct in value, representation and interpretation.

Data accuracy is probably one of the easier quality metrics to identify as it manifests itself with frequency in many operational processes. Accuracy for the most part is quantifiable, different techniques and practices can be used to achieve truthful data. Reverification is one solution heavily utilized where “zero defects” is required, it relies on the manual and labor
intensive recertification technique popular in the critical medical observation field. Olsen (2003) outlined a number of other analytical techniques that can be performed to achieve acceptable accuracy including element and structural analysis, value correlation and value inspection. Each technique led to an acceptable tolerance level that could be applied to attain a reasonable “fitness for use” metric for a specific data purpose. In addition to accuracy techniques, Loshin (2010) provided a number of characteristics of data accuracy including the ability to reference data to a system of record such as a MDM or external certifiable source such as Reuters, Bloomberg or Morningstar. He also emphasized the importance of value definition, precision and acceptance at a domain level and in the case of shared data elevates the requirements to an enterprise level.

Although identifiable and quantifiable, data does have a subjective element when leveraged in different domains and across domains. Maydanchik (2007) pointed this out when describing new or repurposing of data for additional function, citing a defect rate of 15% as being accurate and acceptable for customer addresses in a marketing campaign, but inaccurate and unacceptable for customer billing. Dyché (2006) encouraged the repurpose of the data asset across lines of business, divisions and business processes to maximize its potential value.

The intrinsic characteristics of data highlight an important accuracy consideration where data may be valid, but not accurate. English (1999) provided an example where a date value of 01/01/1901 adheres to a value domain but may not be an accurate representation of a current employee’s birth date. Likewise, Maydanchik (2007) and Olsen (2003) noted the considerable impact of “default” values where blank values are not allowed, providing false or misrepresentation of facts, whereas null values effect determination and aggregation rules.
Data accuracy is in the eye of the beholder, it is affected by domain of use and “the real world” objects it depicts. Complete, 100% accuracy may not be attainable (or desired), but a process of continuous improvement will eliminate defects, identify new features of data and provide controls around the pervasive behavior of data through its lifecycle. Redman (2008) characterized all organizations as being content providers, and additionally in many cases also being content consumers in a world where accurate data is not an added feature but a prerequisite for entry.

2.5.5 Completeness

Each real world object in information terms represents a grouping of datum that provides meaning and ultimately value to an organization. The business object data characteristics consists of mandatory and optional elements that provides a big picture view. Loshin (2010) reinforced this by stating data models should only include relevant data and not include redundant or unnecessary elements. Though sound in theory, the application of such within the financial services industry is complex. Multiple financial instruments exists such as fixed income, equity, derivatives and money markets types and in many cases these business objects reside within a common database table structure. This makes completeness rules challenging as a mandatory attribute for one instrument such as a coupon or issuer may be optional or not applicable for other financial instruments. Maydanchik (2007) stated that data models should be designed along logical model practices and include “identity, cardinality, and inheritance rules,” provisioning for entity identification, sub-typing as well as column and table relationships.

The biggest impacts of incompleteness is the ambiguity that it presents within the process, a reliance of knowledge workers to identify and mitigate risk and as Wand and Wang (1996) stated the threat of operational and business decisions being made on incorrect data.
Initiating completeness checks like most other dimensions starts with identifying the use of data and the underlying business objects being manipulated. Maydanchik (2007) offered a number of completeness checks. At the atomic level incomplete entities may have missing values or populate data that is not applicable. From a lineage perspective, Maydanchik reflected on the interrelationships of objects such as cardinality (important in coupon and dividend schedules) and missing records, a critical element for multi-phase transactions such as a trade and settlement workflows.

Lee and Strong (2003) defined completeness as not missing data and having the breadth and depth to meet the task at hand, they also throw in a cautionary measure that the data must be sufficient to meet our needs which can and will change over time.

2.5.6 Consistency

Information may have different meanings depending upon how data is collected, transformed and ultimately presented, which in turn can impact the creditability of data. McGilvray (2008) defined consistency as an equivalence measure of same data throughout the organization. Loshin (2001) noted that consistency can be “curiously simple or dangerously complex” referring to single domain, single value data sets versus enterprise wide use of a particular data set. Organizations strive to keep a consistent and standard meaning to information as it passes through the enterprise. English (1996) stated that the more pervasive data is within an organization with replication, synchronization and transformation the more difficult it is to keep consistent. English advocated an enterprise wide use of data standards to ensure a common understanding and meaning are applied to both business and technical perspectives.
Inconsistency like incompleteness can exist at both the data element and business object level. Olsen (2003) identified at least three data anomalies that can lead to data inconsistency:

- **Value representations**: For example MA, Mass. and Massachusetts all refer to a US State and each may be correct within their domain.

- **Change inconsistency**: Reference values may be added to a domain that is not considered the authoritative source for the data element.

- **Data entry inconsistency**: Domain knowledge and experience may vary between resources that may lead to inconsistent entry behavior for similar business transactions.

Both English (1996) and Olsen (2003) recognized the impact that inconsistent development standards can have on the underlying data model. English advocated the use of development standards such as singular modules to address update, insert or delete actions at entity, sub-type and what he calls “natural real world events.” While standards and a consistent development model contribute to consistent data behavior, Loshin (2009) and Dyché (2006) encouraged the implementation of MDM and Metadata Repository systems to provide an authoritative source for integration, meaning and perspectives of shared data throughout the organization.

### 2.5.7 Timeliness and Currency

A number of researchers have noted that the fluid nature of data allows it to ebb and flow throughout an enterprise creating accessibility and delivery expectations. Loshin (2010) described timeliness as the measurement of time between when information is expected and when it is made available. McGilvray (2008) referred to this as the “electronic float” but she also
identified an “information float” as the lag between when data changes in a real world object and when the underlying components (and systems) reflect the change. Redman (2001) associated currency with timeliness expressing data can only be up to date as a result of timely information chain.

The on-time delivery of information is classified as a pragmatic quality measurement (English, 1999), that assists in effectiveness and efficiency of the resource, which could be a knowledge worker or automated system. The antonym of this is ineffectiveness. Olsen (2003) described two scenarios where timeliness is examined. The first scenario being a very accurate database that takes several days to release data that is needed immediately. The second a historical warehouse where data is analyzed over a period of years, in this case electronic float and currency issues are not significant. Redman (2008) made the case for historical data timeliness as it consists of the majority of the data in a database and as such is more error prone having had more opportunity to fester and decay.

Wang et al. (1996) described three factors that affect timeliness:

- **System Currency**: Gauges the time delta between real world updates being reflected in its underlying data elements.
- **Volatility**: The frequency of updates to the real world objects.
- **Execution Time**: This reflects the time that the data is needed for process execution.

To identify timeliness, currency and accessibility issues, Redman (2008) advocated the use of audit and timestamps for creation and modifications to assist in measuring all three factors. Timeline constraints can also be imposed, such as continuity and duration rules and advanced duration analysis based on state specific and action specific conditions (Maydanchik, 2007). In
addition to these metrics for internal data, Loshin (2009) supported the inclusion of time clauses, considerations and penalties in any and all contracts (service level agreements) with third party information sources.

**2.5.8 Identification**

The ability to manage data as a corporate asset requires being able to uniquely identify and classify information. Just as other assets such as books (ISBN), consumer goods (bar coding) and even financial assets (Bloomberg mnemonic) can be leveraged to locate and differentiate various real world assets, the information asset should also be just as readily identifiable. The cost and impact of not being able to locate information is critical, knowledge workers incredibly spend up to 35% of their time searching for information and are unsuccessful in 50% of the cases (Redman, 2008). Data, however, driven by its intangible and pervasive characteristics provides an additional challenge to the identification conundrum. The uniqueness issue has become so critical and expensive especially in the customer management and marketing operations that it has spawned a sub-practice called deduplication within the data quality spectrum, specifically to handle duplicates identification, matching, survivorship and merging techniques (Chaki, 2010).

The ability to uniquely identify, store and retrieve data relies heavily on getting the data model correct. Although databases are built on the rules of normalization, functional dependency, uniqueness and referential integrity, many of these traits are built based on the physical model and are susceptible to duplicates. Olsen (2003) offered the case of surrogate keys over natural keys as a prime example and advises placing data elements into categories such as identifiers, descriptors, quantifiers and free-formed text. Maydanchik (2007) pushed the logical model as a design consideration (along with conceptual models) that is often neglected in the race to create the physical structure.
Fisher (2009) promoted the use of proactive quality rules around standardization, geocoding, house holding as well as the use of metadata repositories and MDM systems to provide meaning and authoritative source mapping for key business objects such as customer, product or other business domains. Dyché (2009) also cited the benefits of CDI as standardizing householding, customer identity recognition and duplicate processing into a contained single source, thus reducing process replication and operational costs.

### 2.5.9 Creditability and Trust

Creditability is a paramount component of any business relationship. Kouzes and Posner (2008) defined the term simply as “Do What You Say You Will Do (DWYSYWD).” Creditability earns trust, an invaluable and needed trait in relationships with financial advisors, equipment, resources or data. It is the trust factor and belief in these resources that allow for both strategic and operational decision making (Redman, 1998 and 2008). Many business professions adopt the immortal words of President Ronald Reagan “Trust, but verify” when addressing creditability concerns. Federal organizations such as the FDA (Food and Drug Administration) and USDA (United States Department of Agriculture) provide assurances within their verticals, while professional bodies such as medical boards and bar associations certify the professional credentials we rely for specific services. Yet the data that many of these organizations or professionals rely on may not be up to a standard of quality to meet all expectations.

Once trust is lost, it is almost impossible to regain, an aversion to risk and repeat occurrences push resources to come up with their own remedy. Hakim (2007) pointed to the lack of data creditability as a prime reason for the proliferation of spreadsheets and more recently Microsoft Access databases which he referred to as the “islands of information period and then quotation mark” These siloed islands are cited by the Drucker (1996) as lacking control,
enterprise knowledge, security, governance and risk, providing flawed or incomplete information towards decision making. Non-creditable data and systems can also lead to low morale among employees when forced to use systems they have no belief in and spend time cleaning up data that they are not responsible for (Redman, 1998). For customer facing domains the results can be more devastating, putting a firm’s perception, goodwill and ultimately earnings are at stake, when a client loses faith in the company or the product (English, 2009 and Dyché et al., 2006).

As an organization, there are a number of steps one can take to protect ones integrity. From an information governance perspective audits, continual education and/or certification encourage an environment of trust and accountability, in addition to promoting the severity and zero tolerance policies for violations. Key finding of the Canadian Institute of Health Information (2003) included fostering a culture of quality throughout the organization, continual assessment and prioritizing the types and impacts of data as the basis for risk mitigation and ensuing trust as a data provider.

2.5.10 Data Quality Summary

Wilson (2010) believed that data had begun to gain in significance as a corporate asset necessitating a level of quality that is uniquely identifiable, accurate, complete, concise and relevant, and that meets time and currency constraints. Some industry leaders in the data quality realm have proposed additional data dimensions including presentational (Loshin, 2010), ease of use and transactability (McGilvray, 2008). But many of these supplements can be addressed within the provided dimensions or other governance practices such as metadata and MDM. Further categories may also possibly lead to additional confusion to an already complex framework. Wang (1996) defined higher order categories such intrinsic, accessibility, contextual and representational to house the varied quality dimensions. The higher order categorization may
prove helpful in identifying different root causes of quality issues and may also be used for higher level reporting and training on types of quality defects, but is not a prerequisite when initiating a data quality program.

Data profiling provides another data quality technique that should be given consideration. Profiling uses data analysis techniques to determine statistics around data content. The statistical results for a data set can include pattern analysis, mean, median values as well as outliers, standard deviation and frequency distributions to aid the knowledge worker identify patterns, trends and tolerances in using “a what is versus a what’s supposed to be” rule of thumb. Olsen (2003) referred to this as an “inside in” technique, while dimensions are considered “outside in” investigations of quality.

The quality of data continues to be an Achilles heel for many organizations, yet the steps highlighted ingrained as part of an overarching data governance program can help gain a better understanding of data, its “real world” use, idiosyncrasies and creditability.

2.6 Metadata Management

A business ecosystem is always in a constant state of motion, products are created or retired, employees come and go and IT systems are merged, enhanced and/or replaced. In such a state of flux, companies tend to lose sight of the fact that they run the risk of extensive knowledge loss with every resource shift. Information is no exception, unless it is retained in a documented, understood, classified and retrievable format, the loss can be extensive. Within the data management realm the best and in some cases the only opportunity to attain and preserve this knowledge is through the use of a metadata repository (MDR).
2.6.1 Metadata Definitions

As noted earlier, metadata has been commonly referred to as “data about data.” Wilson (2010) elaborated on the term Meta- as being a prefix coming from Greek origin that defines “beyond, among and along with.” Metadata in its broadest sense is a semantic technology that brings meaning to data. Popular examples include Dublin Core which provides a base fifteen elements that describe a resource such as a document, form or a report (DCMI, 2007). Though basic in structure Dublin Core elements describe pertinent information that may not be directly included in the resource such as title, subject area, creator, publisher and last modified date.

Two key benefits can be realized from metadata. Firstly, metadata is measurable. It can be easily aggregated and reported on. Secondly, it can be layered. Marco (2004) identified six layers (metadata sourcing, integration, repository, management, data mart and delivery). Loshin (2009) also identified layers, but used a more business centric perspective by collecting data elements into hierarchical structures representing real world business entities.

Researchers have been in almost uniform agreement about the benefits of metadata, yet some differences in perspectives exist. Fisher (2009), Dyché et al. (2006), Olsen (2003) and Maydanchik (2007) all cited the benefits of metadata in their teachings and writings. But Loshin (2009) found the term “benign” and not truly representative of the wealth, depth and breadth of information available within the metadata domain. Interestingly, consensus is not achieved with the name metadata either, Redman (2001, 2008) calls it Data Resource Data, English (1999) refers to Data Definition and calls the store an information dictionary, while Dyché et al. (2006) refers to the traditional term metadata.
Disregarding the naming semantics, each researcher recognized the needs, benefits and complexities of implementing a metadata management practice within the broader scope of overall data governance program.

2.6.2 Metadata Components

The composite elements of metadata are small. Dyché et al. (2006) identified three distinct types of metadata: business metadata which is more descriptive in nature, database (or technical) metadata and application metadata which is definition based. Tozer (1999) came to similar conclusions identifying the categories as usage (business), structure (technical) and population (application). Kimball (2008) added a fourth member - process metadata to address the various integrations, meanings and usage of data flow in and out of a data warehouse.

What constitutes metadata is also up for debate. Redman (2001) provided constituent parts of a metadata program as follows: 1) cataloging – used for the collection, meaning, classification and retention of information; 2) data modeling - for designing and visualization of the conceptual, logical and physical layering of data and 3) data standards - to manage the internal and external policy of information.

Loshin (2009) also referenced each of these and believes that all can reside within a unified metadata repository something that Kimball (2008) referred to as the “the holy grail of data warehousing.” To achieve Loshin’s objective, the models and repository would have to be kept in synch using a commonly understood markup language such as OWL (Web Ontology Language).
2.6.3 Metadata Catalog

Cataloging is the term often used for the collection, identification and management of metadata, similar in concept to a library system, metadata catalogs allow for locating data items from a number of perspectives including business alias, descriptions to record and field definitions. As data competency and knowledge data increase, information can be further classified using taxonomies and ontology concepts. Loshin (2009) identifies seven levels of metadata: 1) business, 2) reference 3) service 4) data element metadata 5) business definitions (aka business semantics), 6) information architecture (aka Taxonomy / Ontology) and 7) data governance management (Information Policy). Khatri et al. (2010) provided different naming but was conceptually in agreement.

2.6.4 Business Metadata

Inmon et al. (2007) called business metadata as the means of capturing enterprise knowledge, collecting key business terms, their definitions and their many and varied alias, synonyms, rules and values as it crosses departments and enterprises. Another key responsibility of business metadata is to provide details around data ownership, stewardship and information sharing. Redman (2008) highlighted this contentious issue in stating that information sharing does not come naturally to individuals. Dyché et al., (2006) agreed that data hoarding for fear of accountability is a pervasive problem in most organizations. For information to be leveraged as a corporate asset, it needs to owned, shared, collaborated and communicated on an enterprise level (Inmon et al., 2007), in much the same way different divisions such as sales, marketing, finance and legal would collaborate on the launch of a new product. An important step in this process is making data available, findable, understood and usable which starts with a well defined business metadata collection plan.
2.6.5 Technical Metadata

Technical metadata has been in existence since the dawn of computers (Inmon, 2007). In today’s environment technical metadata is ingrained in and forms the basis of most information centric applications such as databases, ERP systems, ETL platforms and documents. Kimball (2008) lamented the lack of a unified cohesive repository for technical metadata, citing various vendors’ proprietary approach and aversion to sharing their information with others. Technical metadata is critical to gaining an understanding to technical aspects of different systems, and when aligned with its business counterpart can provide an important link between business meaning and technical representation.

2.6.6 Process Metadata

Redman (2008) identified only two distinct states for data, 1) static form - when it resides in a data store 2) motion form - when it is being retrieved, transformed or moved between stores. He believed that it is the latter state that provides value proposition and generates revenue. The movement and transformation of data is facilitated though any number of business processes. The management, monitoring and measurement of these defined processes also produce a rich vein of metadata. The importance of metadata to the process is on par with that of business metadata. Kimball (2008) identified this and asserted the need for a separate category of metadata for business process management. Process metadata can assist both the business understanding and the technical design of a process job. In the current business environment, processes can be as simple as running a report or be extremely complex structures representing various workflows, dependencies and manual interventions. Redman (2008) referred to this as a Small-p versus Big-P with the requisite metadata classifying the different processes. The impact of a one of the two types of processes may have different consequences for the organization.
2.6.7 Key Concepts – Semantics, Ontology and Taxonomies

Everyone within an organization touches and shares data at some point in time, data is copied, parsed, transformed and redistributed continuously. Yet, the complexity and understanding of data is underestimated according to Redman (2001). Knowledge workers in one department may have familiarity with information within their domain, but lack awareness of the lineage or impact that their actions may have on others. To gain an insight into data at enterprise level, a common understanding of business terms, their underlying data elements, lineage and real world usage are imperative. In that context, three concepts can be identified.

- **Semantics**: Provides a richer meaning to common business terms, with drill down capability offering further details of core characteristics of the term. Loshin (2009) provided an example where Baltimore airport and its alias BWI (Baltimore Washington International) can be used interchangeably but only in the context of flying from a specific airport, when used in a different context Baltimore is a city in Maryland.

- **Taxonomy**: Atomic data represents single piece of data in its most granular form such as a cell on a spreadsheet or a field value in a database. It lacks context, classification and real world structure to a point where it has no immediate value. Data only becomes valuable when is organized into one or more real world entities that can be understood and managed. Taylor (2002) referred to the term taxonomy as giving names to things. It is also closely related to classification of elements and can represent a hierarchical class of semantics. For example, taxonomies can represent and organize derivatives as a hierarchical structure with semantic elements and classifications representing the different types of derivatives such as swaps, futures and forwards.
• **Ontology:** Though sometimes used interchangeably with taxonomy, there are distinctly different. Dogcag et al. (2002) defined taxonomy only as representative of class/subclass relationships where ontology provides for more domain enrichment. Sheth (2003) elaborated on this by stating that it is a conceptual representation of the world according to an organization that may include definitions, models and informational relationships. Having an updated enterprise wide ontology with drill through capability to classification groupings and semantic business meaning can provide an enterprise with a 360 degree view of the data those businesses crave but finds difficult to achieve.

2.6.8 **Information Policy**

The information policy of an organization governs the rules, security and internal and external policy as it relates to a set of data elements. In the new era of regulatory reform such as Dodd-Frank, Basel II and III and Sarbanes-Oxley, more emphasis is being put on financial service organization’s information. Loshin (2009), Redman (2008), Fisher (2009) and Dyché et al. (2006) recognized the importance of data policy within a data governance framework in meeting transparency and disclosure requirements. As shown in Figure 2.14, the ability to determine meaning from the single data element, through real world entities and application of metadata and classifications is a critical component of this objective.
2.7 Benefits of Metadata

The benefits of a current and continuous metadata strategy are many and varied. Forrester Research (Agosta, 2002) claimed that metadata management could lead to operational, design, development efficiencies and allow business and technical users’ converse in a common language. Metadata has been called the “DNA for data warehousing” by Kimball (2008) and that the whole warehouse architecture is so dependent on good metadata to be called metadata driven. Maydanchik (2007) called the repository the encyclopedia of knowledge about the data, while Redman (2008) and Inmon et al., (2007) saw it as a means to solve the findability issue that plagues the knowledge worker. A common structure and approach to data management will
result in simplification and efficiency in process and in answering how, where and by whom data is being used, its scope, population and lifecycle (Tozer, 1999 and Redman, 2008).

Metadata also provided meaning to MDM objects. Dyché et al. (2006) and Loshin (2009) both referred to the different interpretations of a customer within an organization, the marketing’s perception of a customer differs greatly from that of the helpdesk or the legal department. Dyché stated it as the one big differentiator between customer data integration and common data integration. Marco (2004) believed that the key to a company’s prosperity is dependent of how it manages its knowledge, with the metadata repository being a large contributor to this success. Future developments such as cloud computing will also further encourage the need for stronger metadata and security, as data integration using web services becomes more prevalent (Merritt, 2008).

2.8 Metadata – Hard to Harness

The benefits derived from metadata should in theory make it a priority for any organization to implement and maintain, yet just the opposite is true. Redman (2008) points to the complexity of keeping the repository current, political motives to hoard data and lack of accountability as the primary reasons metadata initiatives fail. Astonishingly, Maydanchik (2007) claimed never to have seen or implemented a comprehensive metadata warehouse, while Olsen (2003) called metadata management projects a resounding failure due to the time, effort and overall enterprise understanding that knowledge workers’ must possess to effectively fulfill the post.

There are also some inherent risks in using a metadata repository. English (2009) pointed to the fact that without extensive consensus there is no way to determine if the correct context
has been applied to data elements, in addition reconciliation of business terms and data standards is hard to quantify and maintain. A policy such as “data must be of sufficient quality” is an objective metric that depends on context and domain usage.

2.8.1 Metadata Summary

The organizational needs for metadata management are many, yet one could say that “the spirit is willing, but the flesh is weak” when it comes to the continuous improvement aspects of metadata management. Technical metadata is available but mostly in siloed and vendor proprietary state. Business metadata is difficult to cultivate and can initiate political battles over ownership, accountability and sharing. Yet it is the fusion of technical and business metadata into domain specific and domain independent ontology’s inside a unified repository that Kimball (2008) and others finds so compelling.
2.9 Summary

Academic researchers and industry experts have brought different perspectives, methodology and ideas to data governance equation. Yet the one ideal that is evident in the literature is that the pervasive nature of data in the organization necessitates a well planned, iterative and adaptable approach of data governance. Though no platform, no architecture or no vendor tool can meet the data governance needs for any and all organizations, proper planning and implementation of the items addressed in this chapter as part of an overreaching data governance initiative provides a better opportunity to deliver data to the right system, at the right time and with the right meaning intact to meet business requirements.

The next chapter details the additional methodologies used as part of this research. It identifies the qualitative methods used and the adoption of a triangulation approach to correlate the results from the different phases.
Chapter 3 Research Methodology

Chapter 2 provided the first phase of this research. It reviewed the existing academic literature on data governance and how one could initiate and maintain a governance program. This chapter presents the two remaining phases of the research methodology used to test the thesis. These phases are 1) a case study of a data governance program within a major financial services organization, and 2) a questionnaire examining the state of data governance within various banking and asset management organizations.

Holter and Kanneberg’s depiction of the elements of a research as seen in Figure 3.1 (in Stoveland, 2008) has been adapted to demonstrate the reliability and validity of the qualitative techniques used in this thesis. The triangulation approach is used to confirm the case study findings and to provide correlation between the literature review (DG theory), survey (consensus data) and the case study (DG Implementation).

![Figure 3.1- Elements of a Research Project. (Holter and Kanneberg in Stoveland, 2008)](image)

This researcher decided to use a single case study approach to investigate the techniques used by a major financial services organization in implementing their data governance program.
The single source case study was selected over the use of multiple sources due to the complexity of the subject area and vast amount of information to cover. It also allowed for a more detailed investigation and in depth analysis of the impact of data governance within the organization. The approach used qualitative techniques leveraging interviews with senior level information technology (IT) management, in addition to observations and discussions with the operational resources that maintain the program on a daily basis. In circumstances where qualitative findings could be identified and supplemented with quantifiable metrics both sources were utilized.

3.1 Research Methods

From January 5th, 2011 to April 30th, 2011, this researcher undertook a case study research of a major global financial services organization called Pioneer Investments. The asset management company is a subsidiary of the Italian multinational UniCredit Bank. Headquartered in Milan, Pioneer also has operational hubs located in Boston, Dublin, London and Munich. In this study, a number of research methods were used including participant interviews, of which there were 20 interviews in total. In addition to the interviews, observations were made during the many and varied governance and quality meetings that took place during this period. These sessions covered many levels of data management coordination, from the strategic monthly data governance steering committee meetings to the more operationally focused weekly data quality support groups and project management status meetings. Examination of the organizations current Body of Knowledge (BoD) and operational practices was also in scope, covering areas such as issue resolution workflows, database design documentation and data lifecycle and retention policies.
The interview process consisted of interviewing Pioneer Investments executive and senior level IT management both in the United States and Europe. Although other departments participated in the data governance initiative including human resources, finance, investment operations and legal, the majority of the focus, budget and overall impact resided within the information technology domain. The key IT resources who agreed to be interviewed are identified in Table 3.1 by their business titles.

<table>
<thead>
<tr>
<th>Seq. Nbr.</th>
<th>Business Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Global Chief Technology Officer</td>
</tr>
<tr>
<td>2</td>
<td>Chief Technology Officer - United States</td>
</tr>
<tr>
<td>3</td>
<td>Chief Technology Officer - Europe</td>
</tr>
<tr>
<td>4</td>
<td>Global Chief Data Officer</td>
</tr>
<tr>
<td>5</td>
<td>Global head of Software Development</td>
</tr>
<tr>
<td>6</td>
<td>Vice President of Infrastructure</td>
</tr>
<tr>
<td>7</td>
<td>Vice President of Trading Applications</td>
</tr>
<tr>
<td>8</td>
<td>Head of Center of Excellence - Securities</td>
</tr>
<tr>
<td>9</td>
<td>Head of Center of Excellence - Derivatives</td>
</tr>
<tr>
<td>10</td>
<td>Senior Data Architect/Steward - Investments</td>
</tr>
<tr>
<td>11</td>
<td>Director of Enterprise Data Warehouse</td>
</tr>
</tbody>
</table>

Table 3.1 IT Resources Interviewed

The interviews were supplemented by a targeted survey. The objective being to solicit feedback from industry leaders within the financial services industry regarding their data governance practices. The candidates selected represented a broad spectrum of resources including legal, marketing, investment management as well as the IT department. The survey
consisted of ten multiple choice questions all of which were mandatory. The questions were used to identify what the respondents considered a data governance program to entail, as well as DG’s significance, value, status, sponsorship and challenges within the broad financial services industry. The survey can be found in Appendix A.

In July 2011, the survey was created and published to the Google docs website, with requests to participate sent via email to the respective candidates. After a two week period, the survey posting window was considered closed. The results were then aggregated, visualized and analyzed.

3.2 Summary

The primary objective of using the two different research methods is to gain both a deep and a broad understanding of data governance principles and practices within the financial services industry. The case study of Pioneer Investments provides an in-depth view of a single data governance program, its organization, adaption and operation. Although the survey lacks depth, it provides a broader consensus around how others in the industry perceive data governance. Using this approach, this researcher ensured that the issues, findings and approaches used by Pioneer Investments can be trusted, was systemic and correlated well the broader goals and objectives for data management throughout the industry as a whole.

The next chapter looks at the findings from both of these research methods.
Chapter 4 Findings

This chapter examines the results of the survey. It is divided into two parts; the first examines the results of the case study of Pioneer Investment’s data governance program. The second analyses the results of a survey questionnaire completed by a senior level management within the financial services industry.

4.1 Case Study

The case study survey goal being, that through investigational proceedings of the following components, aligned with observations of their daily use will prove the hypothesis as described in the thesis statement. The work sought to facilitate understanding of data governance in the following categories.

- **The data governance model**: Outline the model for data governance within an asset management organization, their components and their effective implementation.

- **The support framework**: How Pioneer Investments manages the daily data requirements and integrity of their data. The introduction of an enterprise information management (EIM) and centers of excellence (COE), their organizational makeup, resources, ownership and workflows.

- **The value proposition**: The cost of the Pioneer Investment’s implementation remains confidential, thus making a dollar value return on investment (ROI) impossible to gauge. There are, however, a number of other qualitative measures vindicating the necessity and decision to implement and maintain a strong data governance program.
4.1.1 Background

Pioneer Investments is a global asset management company with assets under management (AUM) of €185 billion (2010). The company specializes in the management of mutual funds and institutional accounts using a global platform for the purchase and sale of securities and/or commodities. Up to 2004, each international location acted semi-autonomously having separate trading, reporting systems and processes. As a result, it was almost impossible to collect, identify and report critical information to meet internal and external mandates.

- **Exposure:** A number of significant events such as Bernard Madoff’s $50 billion Ponzi scheme (2008), the Irish banks meltdown and the threatened insolvency of the US car industry led to many financial services companies scrambling to identify their risk exposure to these and other events.

- **Reconciliation:** The globalization of stock exchanges has allowed large multinational organizations gain listings on multiple exchanges. The non standardization of security identifiers across exchanges has led to the use of different identification protocols to represent the same underlying equity of the company in different regions of the globe.

- **Meaning:** In recent years, there has been an explosion of new and more complex derivative trading instruments such as CDS (credit default swaps), CDO (collateral debt obligations) and CLO (collateralized loan obligation) whose makeup and meaning are confusing to most, even to those who work in the industry.

- **Regulation:** The financial services industry has always been a highly regulated area. The globalization of asset management creates even more regulatory requirements that must be adhered to.
Pioneer Investments took a proactive approach to its governance plans, identifying early on the risks of exposure and regulatory non-compliance. As a result a strategic plan of action named GISP (Global Investment Systems Platform) was initiated to better manage data at the enterprise level. This investment required an enormous shift in platforms, resources, collaboration and most importantly a shift in the organizations perceptions, value and mechanisms for integrating information.

### 4.2 Organizational Structure

In 2004, Pioneer’s data management resources, in a similar vein to the platforms they managed, were geographically dispersed with little or no interaction with their overseas counterparts. IT management structures were application specific with little regard for the data outside their application domain. GISP was about to change this, leading to a number of appointments and organization structure changes within the IT group. The first was the appointment of a Global Chief Technology Officer (GCTO) to establish and oversee the use of technology and resources on a global basis. The GCTO removed the siloed approach within the regional IT groups and moved towards a more unified plan for software and hardware purchases and utilization. It also provisioned a ‘global pool’ of IT resources within a Global IT management structure, the availability and utilization of which is managed at an enterprise rather than a regional level.

The second appointment was the position of Global Chief Data Officer (GCDO), a title and position that is relatively new within both the IT and the financial services industry. The GCDO’s responsibility is the collection, organization and management of all the company’s investment data from inception to disposition. This ownership of data transcends across all
sources and all application platforms both internally and externally. It ensures that the data is of acceptable quality, used correctly and within compliance terms by the organization. The position is also accountable for contracting, licensing and integrating of all external third party data that Pioneer Investments purchases such as Standard and Poor, Bloomberg and Reuters ratings, indices and benchmark information.

The importance and relevance of the Chief Data Officers position and its location within the organization chart in Figure 4.1 is no coincidence. The CDO is strategically positioned in alignment with the CTOs for the United States and Europe, reporting directly to the Global CTO. This gives significance and credibility to the critical tasks and responsibilities that the position holds.

**Figure 4.1 Higher order organizational structure for IT department**

The higher order chart displayed above can be further broken down to show the makeup of the regional data management groups. Although the data management group legally falls under the management of the IT organization, there is active participation from other departments and resources. Figure 4.2 depicts this function as a combination of both functional and technical expertise that is required to manage data.
4.3 Data Quality

The application of data quality techniques before GISP varied across systems with little or no accountability for the integrity of the data. In most cases, quality checks were limited to validations delivered by third party systems, relational database constraints and possibly some lookup or basic validations as part of an ETL process. The introduction of MDM as a component of GISP changed this paradigm. The new MDM systems necessitated the collection and publication of clean, correct, concise and consistent data to all subscribing downstream systems. The failure to supply and sustain an acceptable quality level would lead to system distrust and possibly a move back to the siloed approaches that MDM is supposed to eliminate.
To ensure a continuous goal of quality improvement the data management team identified three maxims upon which their data quality strategy was based:

- **Maxim 1: Data quality (DQ) is everyone’s responsibility**: All quality rules, definitions and exceptions are made available to all resources in the organization. Users have the ability to see any and all DQ rules that are in place for any global application owned or licensed by Pioneer. All resources are free to submit DQ rule candidates for any application where they feel current validations are inadequate.

- **Maxim 2: Unified platform for data quality**: Data quality is both fluid and pervasive in its utilization and as such requires a centralized repository to track the creation, evaluation and maintenance of all definitions enterprise wide. Previously, different groups at Pioneer used various mechanisms for managing data quality definitions, some used word documents, others developed wiki or web pages, while others still “just knew” from experience. GISP introduced a single unified approach to indentifying, entering and maintaining DQ rules regardless of their origination. This was an important and in some cases a contentious step when one considers the two main sources of DQ rules namely business as usual (BAU) and project initiatives (Software Development Life Cycle - SDLC) have different workflows. In the case of project work, many DQ rules are buried in functional specification, whereas for BAU they were sourced and managed from an operational manual. Pioneer found it necessary to adjust both their SDLC and BAU processes, along with change management to incorporate a common workflow for sourcing, referencing and ongoing evolvement of DQ rule definitions.
• **Maxim 3: Simple utilization:** Here the process for entering and managing data quality rules has to be as simple as possible with minimal intrusiveness into current workflows. The DQ repository is first and foremost a business facing platform and is delivered using a business dialect. It purposely omits any and all technical references to SQL, design patterns or programming code. Pioneer took the additional step of having a graphic designer assist in the graphical user interface (GUI) and usability of the repository to ensure an intuitive and appealing presentation for end users. The system also uses single sign-on authentication allowing all business a user’s seamless and immediate access to the system.

### 4.3.1 Data Management Framework (DMF)

The data management group at Pioneer used these three rules as a basis for introducing a new DQ Platform to the enterprise. After an extensive and unsuccessful search of external providers of data management software, Pioneer made the decision to build their own application - The Data Management Framework (DMF). The motivation behind DMF was to unify the collection, reference and management of data quality for all critical applications in a single web based application. The DMF repository is made readily accessible to the whole organization, without the need for permissions, authorizations or access workflows. This ensured that all users have the requisite information available to them and they have the capability and empowerment to impact data management positively (Maxim 1). DMF has now become the standard functional specification repository for DQ rules identified through SDLC projects. The change to the generally accepted project management principals allowed all DQ rules to be located in DMF regardless of origination. Figure 4.3 highlights these different workflows. Rules can be identified as part of an SDLC approved project, BAU operations or a commissioned DQ workshop. The
DQ rule candidates are then entered into DMF and mapped to specific applications for review/approval by the application owner regardless of source. Consequently rules can be evaluated, approved and coded using the appropriate integration tool leveraging a trained “global pool” of developers (Maxim 2).

The third and final maxim was achieved by making the definition, submission and referencing of DQ rules as facile as possible. To enter a new rule candidate, the user has to complete four simple questions:

- **What can go wrong:** Provide a brief description of risk associated with the DQ rule.
- **What is the correct state:** Provide a description of the correct position/approach.
- **How do we measure:** Provide the measurement between correct and incorrect states.
- **How do we resolve:** Provide the resolution workflow to fix the situation.

Upon completion of the questions, the user selects the application(s) that the rule pertains to and submits the request. Upon submission the rule is assigned a global reference number (DMFID) that uniquely identifies the rule throughout the enterprise. The total time to enter a rule averages between two and five minutes from start to finish (Maxim 3).

All published rules and rule candidates are made available through DMF (and an accompanying RSS Feed) to the whole enterprise providing the capability for any resource to evaluate and/or collaborate on. Each DQ rule candidate has to be approved by the IT owner of the application(s) that the rule is designate for use. From a business user perspective there is now the capability to evaluate the level of quality of the data upon which they are basing their business decisions. This level of transparency did not exist before, but since the introduction of
DMF the business users have become more engaged in understanding IT’s function in the facilitation of data quality and their responsibility in the information lifecycle.

Figure 4.3 New DQ rule workflow

**Exception Management**

The success of DMF as a repository for DQ rule definitions brought about a second use for the repository, as a store for the collection and management of the exceptions produced as a result of the DQ Rule validations. Traditionally Pioneer’s approach to exception handling revolved around the sending email chains between specific support groups. This proved to be ineffective from at least three perspectives:
• **Notification:** Outside of the specific application domain no other resources are made aware of the quality issues. Emails also tend to be cryptic, open to interpretation and may not explain the issue in detail. This inevitably sends the user searching for documentation that may not exist or may be outdated.

• **Maintenance:** Employee turnover in certain support roles is high, adding and removing users from email distribution lists is time consuming. Pioneer also manages email distribution lists through the regional helpdesk, yet the user support role is global in nature.

• **Workflow:** Management may be made aware of the issues through email, but there is no lifecycle workflow to identify the resolution, resources and effort expended.

The approach taken with DMF is to gather all data related exceptions regardless of application or process, and house them in a centralized repository / monitor. From an operations standpoint, DMF is used to reference, assign ownership, monitor work effort and eventually close exceptions using a consistent and managed workflow. The strategic value of DMF produces a set of metrics that provide an aggregated view of the types of errors Pioneer experiences. It also exposes the areas with the highest frequency and volume of defective data, which in turn can be broken down across different domains such as regional, application or data quality dimension. These multiple dimensional views of DQ exceptions allow Pioneer an additional perspective into the root cause of some DQ exceptions that would not be possible when using a siloed or email notification approach to data quality.

The technological approach used in DMF in accepting exceptions from different sources is based on a SOAP web services applications programming interface. SOAP (Simple Object
Access Protocol) is an open standard that is supported by the majority of applications and processes in use in the financial services today. As shown in Figure 4.4, using a set of predefined messages each source can communicate with the DMF web service and route exceptions to the repository. DMF has the built-in flexibility support to accept and render a variable number of fields delivered as part of an exception. This is necessitated due to the dynamic nature of the different sources, exceptions and supporting information required for exception resolution. The only mandatory requirement made of DMF is that the DMFID (Unique DQ Rule ID) is passed as part of the exception message. This facilitates the incoming exceptions being linked to the underlying DMF rule definition, predefined support groups and subsequently the attached resolutions. Exceptions are then assigned to an individuals or support groups who are responsible for exception status, case notes and ultimately case closure once the correct resolution has been applied.

![Figure 4.4 DMF exception integration](image)

Return on Investment (ROI) is one of the fundamental criteria of most initiatives in financial services. Although this metric is difficult to quantify within the data management realm, the ability to identify costs associated with bad or defective data is one that Pioneer values. For the first time Pioneer is able to apply a cost to the exceptions based on an
approximate resource cost and both the lifecycle and the frequency of exception occurrence. Table 4.1 illustrates this fact by firstly assigning each support group with a blended resource cost. The DQ rule is then tracked using creation and closing times to identify the exception timeline to closure. The resulting effort is then tracked and multiplied by the hourly blended cost. The individual costs may be small in the example below, however when case frequency is applied the costs are quite compelling.

<table>
<thead>
<tr>
<th>Support Group</th>
<th>Blended Cost $USD</th>
<th>DQ Rule</th>
<th>#Exceptions</th>
<th>Creation</th>
<th>Closed</th>
<th>Effort</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security Master Group</td>
<td>50</td>
<td>Duplicate Securities</td>
<td>4</td>
<td>9:00 AM</td>
<td>12:00 PM</td>
<td>3</td>
<td>150</td>
</tr>
<tr>
<td>Warehouse Group</td>
<td>150</td>
<td>Missing Holdings</td>
<td>400</td>
<td>9:00 AM</td>
<td>4:30 PM</td>
<td>7.5</td>
<td>1125</td>
</tr>
</tbody>
</table>

Table 4.1 Exception cost analysis

4.3.2 Quality Summary

In the two years that Pioneer has centralized the management of data quality, the organization has become more aware than ever the impact defective data has on the organization. The concept that data quality is an ongoing program rather than an initiative is also gaining broad acceptance. The business users, although slow in accepting responsibility for the ownership role of data, accepted the fact that they must play an important part in ensuring that the data Pioneer uses is “fit for purpose” on an ongoing basis.

4.4 Metadata Management

Pioneer’s investment commitment to MDM provided an opportunity to centrally locate, control and monitor its critical data systems. Golden copies of business critical information are
now secured, approved and published from a single master source. MDM further ensures both the integrity, quality and the delivery of the data to many downstream systems. However knowing which systems integrate with MDM, identifying the content and processes that transform and route the information to the targeted applications remained elusive. Pioneer needed the capability to track applications, processes and data elements and their inter-relationships in order to better understand their business meaning, lineage and the impact each can have at an enterprise level.

With DMF already in place and generally well received, Pioneer decided not to introduce a second data management tool into the equation. Though initially used exclusively for the establishment and monitoring of data quality, DMF has been extended to include a metadata repository to track the relationships between applications, processes and business terms. To highlight the importance of business level metadata over its technical counterpart, Pioneer chose the name business glossary over data dictionary as a more apt name to represent its metadata repository.

4.4.1 Business Glossary

"Many of the truths we cling to depend greatly on our own point of view “

- Obi-Wan Kenobi (Lucas, 1983)

The primary purpose of business glossary (BG) is a business facing repository of artifacts and their associated relationships. An artifact in this context represents a business term, application or process and all encompassing metadata attributes. Artifacts and their apparent “truths” may differ considerably throughout an organization based on their use and specific business context. For example, a marketing division may refer to a stock ticker term as a security
identifier, while the investment operations group may refer to the same identifier as a SEDOL, ISIN or CUSIP during a trade operation. Each department may have a different semantic meaning and context for the information, yet the data value and source of the information may be identical. This can lead to confusion during inter-departmental collaboration. The Pioneer approach was to first and foremost take a business centric approach to the common business terms within the organizational domains. Secondly, the organization build a consensus around the accepted enterprise name for the term and finally enrich with related ownership, artifacts and metadata to build a 360 degree view of the artifact.

**Figure 4.5 A Business Term and its underlying metadata**

Figure 4.5 depicts a typical asset management business artifact named Security Identifier. Each artifact has associated metadata attributes attached that can represent three distinct collections of critical information to the organization.

- **Business Metadata:** The business metadata represents items that are important to the business. It includes alias and synonyms used internally by different business
domains. In addition it also stores external references to business terms as defined by different governing bodies such as EDM Council, FpML (Financial Products Markup Language) or regulatory bodies such as SEC (Securities Exchange Commission) are also maintained by the assigned business owner(s).

- **Technical Metadata**: These attributes provide the technical representations of how IT sees and interacts with the business artifact. Technical metadata for a business term includes underlying database structures such as schemas, tables and fields. The servers, IP addresses and application architecture are also viable candidates for collection. While the process artifact contains metadata for parameter definitions, processing tools, scheduling and other process dependencies.

- **GRC Metadata**: Governance, Risk and Compliance (GRC) attributes collect, maintain and measure compliance based on underlying metadata classifications. Metadata surrounding classification, privacy and retention of artifacts are critical for business continuity, regulatory reporting and e-Discovery litigation. Corporate GRC are responsible for the metadata and usually collaborate with the business to ensure that the data is maintained to applicable standards.

Pioneer’s business glossary currently holds three distinct types of artifacts: applications, processes and business terms. Each is important from an individual perspective, yet it is the relationships between these artifacts that answers the most critical questions posed to both the business and IT groups. The relationships of applications to processes and business terms provide visibility into the sourcing, integration and routing of information. It also assists in identifying downstream impact and root cause analysis as well as reducing the risks involved
with the possible misinterpretation of information. Figure 4.6 illustrates these inter-relationships of the different artifacts. In the example below a central business term can be linked to one or more applications, processes, and data quality rules. A business term can even map to other business terms creating a derived or alternatively a dependent relationship between the artifacts.

![Diagram showing inter-relationships between business terms, processes, applications, and data quality](image)

**Figure 4.6 Business glossary lineage**

The importance of these inter-relationships should not be underestimated. Figure 4.7 illustrates an example where an inaccurate currency value is provided by a third party vendor. The value having been loaded into the trading system is then utilized in order execution, submitted to the financial reporting system and reported to the SEC as part of the NAV (Net Asset Value) for a fund. The business term is also used along with the local currency share price and number of shares to calculate the market value in Euro. The example shows that a single defect within one small application domain can have serious impacts to downstream systems and indeed to the integrity of any associated business terms.
This scenario depicts a small root cause and impact related to defective data. Large financial organizations with multiple systems, processes and domains of integrated information may have multiple levels of lineage to inspect in an acute timeframe in order to limit or fix the damage caused by a single piece of defective data.

Pioneer’s data management workflow in this situation provides a systematic approach, leveraging both DQ rules for exception identification, while using metadata for reference and collaboration. Once a support team is made aware of the DQ exception, the glossary is referenced to identify all lineages that the affected business term may have with other artifacts. The owners of the related artifacts are then consulted about the effective action to take. This can
vary by artifact and may include manual remediation or the suspension of certain processes until the issue has been fully resolved. The benefit of using DMF (Data Quality and Glossary) is its open availability to the whole organization. Each Pioneer resource regardless of their domain can access the platform and view from a business perspective any outstanding exceptions and the impact they may have on any related systems or processes.

4.4.2 Metadata Summary

Metadata management is one of the more complex components of the data governance and one of the hardest to keep current. It requires a deep and varied knowledge of the different types of metadata available, as well as the skill set to stitch, model and classify a vast amounts of enterprise information. Yet the value the business glossary repository brings to Pioneer cannot be underestimated. It removes ambiguity, documents the responsibility of data ownership and presents the informational facts to all domains within the organization. This aligns with Pioneer’s philosophy that data is an enterprise level asset, and that to gain its maximum value it has to be presented, used and maintained by the enterprise as a whole.

4.5 Master Data Management

Pioneer Investments global presence provides it the opportunity to trade securities and manage mutual funds across a geographically dispersed group of locations. Historically, each trading desk performed in a manner of almost total autonomy to the other regions. Each had the capability to add and trade securities within their local trading and portfolio systems with minimal oversight and limited exposure to the larger organization. Reporting was also a challenge as various reports were created of siloed systems, typically using different file formats
and routed using a selection of mechanisms such as email, ftp and file shares to Pioneer’s Milan headquarters for consolidated reporting.

The data issues that manifested at a regional level grew exponentially when the information was relayed to headquarters. Aggregation of data became a manual chore requiring many of hours of mapping and reconciliation, leading to increased labor costs and high resource turnover. The biggest issue however was the lack of transparency into the data and the operations of each division. Various regional and global policies require visibility into an asset manager’s funds to assess holdings caps (SEC Schedule 13g), maximum sales charges (NASD 625) and risk exposure to certain assets classes. The latter highlighted by the credit default swap crisis of 2008 and the escalation in oil and other commodity prices in 2010.

In 2004, Pioneer Investments as a part of the GISP initiative initiated a project identifying the critical investments systems within the organization and for each, to provide a single master data management repository. Each MDM system’s objective is to harmonize the identification, management and distribution of the financial information within its domain, and by holding the golden records becoming the single authoritative source for its responsible domain. Pioneer identified and implemented three distinct MDM systems:

- **Equity and Commodities**: Pioneer’s aim was to have one centralized hub for both equity and derivative securities. The complexity however in the makeup and the diversity of each instrument necessitated the implementation of two separate specialized vendor products to compose the financial instrument MDM layer.

- **Product Management**: Stocks, bonds and other financial instruments make up the component layer upon which mutual and institutional funds are composed. Once
established these financial products themselves need to be managed with authority. Fund products can have their own complexities such as legal and managerial ownership of the entity, local and corporate governance as well as constraints on how and where the product can be sold. It’s important therefore for Pioneer to have consistent, competitive and desirable products to offer their clients worldwide. Pioneer’s product MDM system provides the capability to centrally manage the lifecycle of their full suite of financial products, assign ownership and enforce corporate policy and governance globally.

- **Sales Management:** In the global market place, the sale of financial products has become a complex mapping of channels, regions and clients. Funds can be bought by individuals, institutions, pension or other mutual funds. Each region may have a different setups for commissions, profit-sharing and fees. Pioneer’s sales MDM system keeps track of the different sales channels as well as the revenue, expenditure and allocations needed, this allows Pioneer to tailor sales of products to meet localization needs while managing the financial reporting and profitability milestones set out at a corporate level.

The backbone of all of Pioneer’s MDM systems is the ability to have a single authoritative source for domain information. This allows for a centralized application of DQ standards and for the distribution of authorized data to downstream systems in a consistent and timely manner. To accomplish this data integration component, Pioneer implemented a real time enterprise service bus (ESB) for publications from all MDM systems. Figure 4.8 illustrates the integration flows for Pioneer’s security master MDM system.
Figure 4.8 Pioneer's MDM overview

Each subscribing system is fed consistent securities data in real time from the MDM system once the security has been approved for publication. The central MDM system stores assigns and tracks each instrument with a globally unique PUID (Pioneer Unique identifier). In addition the security is passed with a full suite of security identifiers for instrument identification and resolution globally. For example, the shares of Coca Cola (Ticker KO) may trade using its ISIN (us1912161007), SEDOL ID (B0ZKV12) or TIDM symbol (CCA) based on the exchange and region where the security is traded. This allows for easier aggregations, reconciliations and reporting of the same security across multiple exchanges. In a similar manner to the security master MDM, both the product and sales MDM systems uniquely identify and publish approved
real-time information using a common message format to all their subscribing downstream systems.

4.5.1 MDM Summary

The transition to MDM has provided Pioneer a more controlled environment to ensure the integrity of its critical domain data. This, aligned with real-time data integration allows each regions/system the autonomy to continue to perform their core business tasks. The corporate office now has the authority, transparency and governance functions it needs to manage to a central model for the betterment of the organization as a whole.

4.6 Survey

The second phase of the research consisted of an online survey to gauge the interest and participation of various financial services companies other than Pioneer to gauge existence and interest in data governance programs. The targeted audience for the questionnaire was senior level management both inside and outside the selected organization’s IT domain. The survey was constructed and distributed online using the survey feature of Google Docs. The total number of questions was limited to ten to lessen the time burden and further encourage participation. The majority of the questions consisted of radio buttons for answers that required a single response or selection boxes where multiple selection of options was a valid response. A free form text field was also provided on some questions but was not used by any of the respondents. Participants were guaranteed anonymity so no individuals or business names are included in the results. The survey, which started on July 5th, was closed to respondents on July 25th, 2011 with a total of 34 responses. The results were aggregated, and charted to investigate correlation (and possible
deviations) with both the literature review best practices and the Pioneer Investments data governance undertaking. The survey can be found in Appendix A.

4.6.1 What is the Status of your DG initiative?

Question 1 of the survey sought to the current state of data governance within their organizations. A combined total of 71% of respondents were either in production of the testing phase of a DG initiative. This result demonstrates that financial services companies were aware of the need for governance and in most cases had taken steps to implement a program.

![Figure 4.9 Data governance program status](image)

4.6.2 Data Governance Competencies

Figure 4.10 shows the results when respondents were asked identify what components they felt comprised a data governance framework. Data quality, integration and MDM were identified as the main constituents of most programs. Metadata, security and lifecycle management were all seen of equal importance. The fact that metadata was required for classification and feeds into both security and lifecycle confirms this researchers point that metadata is often misunderstood and difficult to harness and maintain. Database design was not seen as being as significant, yet bad database design can lead to duplicate data and additional redundant stores of information.
4.6.3 Data Governance is a function of which group(s)

Figure 4.11 provided evidence that the ownership and accountability for the data governance program could be shared between IT and business. Of all responses, 71.4% considered the function a shared responsibility. This was consistent with the literature review in Chapter 2. Interestingly, an additional 14% saw it as solely a business function. This showed a trend that organizations now see data ownership outside the domain of IT.
4.6.4 Data Governance Drivers

Figure 4.12 showed that compliance and risk are significant drivers for data governance. An interesting second driver was business need which ranked higher than both cost and IT needs. The results show that additional drivers for DG besides GRC have emerged and that the business users have begun to understand the benefit of effective data management.

![Figure 4.12 Data governance drivers](image)

4.6.5 What is the scope of the DG initiative?

The survey questioned the scope of the data governance initiative. As Figure 4.13 shows a resounding 72% were running governance enterprise wide. This is consistent with organizations that have a strategic view of DG program. Project based and departmental initiatives are sometimes used as part of proof of concept where executive buy in may not initially be sought or received. Smaller targeted initiatives around specific areas of operation may also account for the narrow approach and scope.
4.6.6 Obstacles to Data Governance

What were the major hurdles affecting the implementation of data governance programs? Figure 4.14 shows that executive buy-in with 19% of the responses was seen as the biggest obstacle. The remaining responses were evenly spread among the remaining options. The results can be interpreted to mean that there are many obstacles of equal importance that may hinder a DG initiative. The results reflected many of the barriers raised during the literature review and those experienced by Pioneer Investments in their program. Interestingly return on investment was not seen as an obstacle, which may show that financial organizations have realized the benefits of data governance.
4.6.7 Executive participation in a DG program should be positioned

What is the optimal timeframe to have executive participation in a DG program? Both the literature review and the case study recommend executive sponsorship from program infancy. Figure 4.15 showed that 71% of the respondents considered executive buy-in from the start as being the most beneficial. The 29% that recommend partial participation correlated well to the combined departmental and project based DG scopes (combined total of 28%).

Figure 4.15 Executive participation in DG

4.6.8 Data Governance Planning

The results from Figure 4.16 tend to support the thesis of this study that for data governance to be effective, it must be productionalized and pervasive within the organization. Many companies have governance as part of their project planning process, yet only eight (21%) respondents identified governance as part of operational support and three (8%) have DG integrated with change management. This is a deviation from Pioneer Investments approach, but this may be due to the lack of maturity of many organizations DG program. Interpreting the responses, project planning may lend itself well as a starting point. In many data centric projects there are usually many non- or semi-formal data policies and validations enacted that may lay the
foundation for governance. As projects become productionalized governance can also follow through change management and into operational support.

![Figure 4.16 Data governance planning](image)

4.6.9 Data Governance Structures

The framework for data governance calls for a number of structures to manage information and quality at a various levels. The existence of a steering committee in 45% of respondents highlights the need for a formalized structure for continuous review as presented in Figure 4.17. In addition, steering committees usually consist of senior level management and/or executive management focused on the strategic planning for data initiatives. Operational support is catered through established data quality meetings. The 22% that do not have a formalized collaboration structure may represent the organizations that are testing or are evaluating DG initiatives. The combined results for steering and DQ meetings (67%) correlated well with Pioneer Investment’s approach with formalized strategic and operational meetings.
Chapter 1 provided many reasons why financial organizations need to better manage their information. The results in Figure 4.18 highlighted the fact that many companies are aware of these requirements and saw the importance and focus on information governance increasing (57%) or at the very least remaining the same (43%). No respondent saw DG decreasing in importance in the next two years. In the case of Pioneer Investments, these results were consistent with their strategic plans and their information platforms. These responses were also in line with the opinions of many of the industry leaders such as Loshin, Dyche and Redman as discussed in the literature review.
4.7 Summary

This chapter reached two major findings:

- **Case Study:** The GISP projects were officially completed in 2009 and have now transitioned into an operational process and support program. Yet, its impact to the organizational and support of information within Pioneer is very much in evidence. There is now a more enterprise level emphasis placed the acquisition, management and disposition of information. From its initial focus of implementing and managing the MDM, data quality and metadata management systems, the IT department has progressed to become an advisor and facilitator of data governance and architecture. Ownership of data has successfully transitioned from IT to the business domains who acknowledged the necessity of their own participation in the data management program.

- **Survey:** The survey as a whole highlighted the awareness of data governance’s importance with the majority of respondents having initiatives in production. The findings also acknowledged the fact that the responsibility and scope for governance does not rest solely within IT, but belongs to the enterprise. There was still some confusion over what components that made up a data governance program and how and where to implement DG. However, the main areas of data quality, MDM and integration were recognized as core and their inclusion in the project planning and SDLC function might signify an initial move towards adaption. Overall, the survey results support the thesis statement: The exponential growth of data, aligned with its pervasive nature into every facet of a
financial services organizational domain necessitates the implementation of robust
data governance program to manage data’s lifecycle from its inception to its
ultimate disposal.
Chapter 5 Conclusions

Data governance has been shown to be both a complex and an evolving competency that has begun to be recognized for being more than a compliance directive or cost overhead. Data volume, types and use look to continue the growth of the last decade. The research carried out in this paper has shown governance is definable and can be successfully implemented and supported. The analysis of Pioneer Investments practices detailed the many of issues faced before governance and how these were remedied by the program. This awareness and the steps taken by Pioneer were vindicated by the survey results as being correct and representative of the scope and actions being taken by the financial services industry as a whole. Returning to the thesis statement, is data governance necessary within the financial services industry? The body of evidence as a result of the research in this paper considers the data governance function a requisite for financial services organizations to implement, conversely its omission can decrease competitive advantage and expose the company to unnecessary risks both internally and with external bodies.

5.1 Contributions

This research has identified a number of areas where additional focus can be applied to data governance Body of Knowledge (BoK).

- Automated Metadata Collection: The collection and use of metadata within the governance program in many cases is overlooked. There are many reasons for this. These include the lack of knowledge and awareness of the rich vein of information that metadata provides. It is often perceived as invisible or peripheral to the more available content that resides in the various databases or documents that metadata describes. The
various volumes, types and classes of metadata that needs to be collected may also seem daunting encompassing technical, business and regulatory policy. Additionally the lineage and stitching required to align the various pieces of data into a logical and cohesive model that accurately describes and measures the underlying business information is also a complex task that requires a high degree of domain and technical expertise. Yet the benefits of metadata are significant. Metadata is a prerequisite for any form of classification, which in turn feeds security and retention policies needed for lifecycle management. Most databases and enterprise applications (such as SAP and Oracle Financials) have metadata layers that can be harvested, matched and aligned to business terms in an automated or semi automated fashion. The can reduce the monotony of managing metadata and can lead to broader acceptance within the organization.

- **Operationlize the governance function:** Although there are some benefits from spot fixes or project tasks to improve the integrity of information, these provide short and unsustainable gains. Redman (2008) identified data in motion as the creator of value. Since data is being integrated and transformed constantly as part of day to day operations, governance too should be incorporated into the process. This includes daily operations, change management and project planning tasks, each are interrelated and the lack of observance in one can have an adverse affect on the others.

- **Collaboration and socialization of the governance function:** Data is a multidimensional object that can be transformed, transferred, searched or aggregated. Its use within one domain may be unknown by others until certain “data events” occur. The transition of the governance role to the business calls for closer collaboration on the
definitions, usage and quality of information being distributed. The ability to map technical terms to business terms (Glossary) and have a discussion forum or blog to converse on data content, context and impacts can prove an invaluable tool for issue resolution as well as a training and knowledge management tool.

5.2 Future research

The research carried out for this thesis concentrated mainly on the management of structured data within the confines of an internal data center. The face of computing and its delivery platform is changing rapidly thus providing opportunities for further research. These areas may include:

- **Cloud computing**: The ability of financial services companies to have public and private clouds allows for the distribution of information outside the confines of the organization. Although the benefits of cloud computing are many they do bring concerns regarding data security and integration needs.

- **Web services**: The decoupling of data from systems has evolved over the years from RPC, CORBA to web based services and Service Oriented Architectures (SOA). Should a governance platform be capable of tracking these web services and the data that they expose, especially those that manifest the data to outside users?

- **Business intelligence**: Analytics has become an important area in financial services, especially in quantitative research. Investing time into the impact, risks and costs associated with dirty data in this critical area can provide justification if any is needed to the implementation of a governance program.
• **The social network:** Facebook, Twitter, LinkedIn and other social media networks have forced many to reconsider what we term as information. An interesting area of governance research may consider how this information is harvested, identified and integrated with internal sources to provide value.

• **Other media:** In addition to social content, there are additional media formats that companies need to manage such as video, pictures, music and other informational and intellectual content. Incorporating a comprehensive content management system (CMS) into the governance framework can provide a full 360 degree view of the corporation’s information assets.

**Other industries:** The systemic use and growth of data can be seen in other industries such as healthcare, utilities, oil and transportation. Each industry may have a different perspective on information and its use, but the overriding need to effectively manage information is a common trait.

The results of this research generally support the advantages of data governance and how it has been applied in a case study in the financial services industry. But competition among a broader range of industries and technologies would, in the long run, make for a more valuable contribution to the literature on data governance.
Appendix A Survey Form

Data Governance within the Financial Services Industry

The survey consists of ten questions and should take an average of just three minutes to complete. The survey will be used as part of my thesis research at Regis University into Data Governance/Management practices within the financial services industry. Thank you for taking the time to participate.

* Required

What is the status of your DG initiative? *

- [ ] In production
- [ ] Piloting
- [ ] Testing
- [ ] Evaluating

Is DG considered a function of *

- [ ] IT
- [ ] Business
- [ ] Both
- [ ] Other: [ ]

DG competencies consist of *
☐ Master Data Management (MDM)

☐ Data Quality program (DQ)

☐ Metadata Repository (MDR)

☐ Data Integration Planning

☐ Database Design/Architecture Planning

☐ Data Security, Classification and Life-cycle policy

☐ Other: [ ]

The primary driver for DG is *

☐ Compliance and Regulatory requirements

☐ IT Requirements

☐ Business needs

☐ Cost savings

☐ Reduce Risk

☐ Other: [ ]
What is the scope of the DG initiative? *

☐ Enterprise
☐ Departmental
☐ Regional
☐ Project-Based
☐ Other: □□□□

The obstacles to DG are *

☐ Getting executive buy in
☐ Budget
☐ Return on Investment (ROI)
☐ Complexity
☐ Business users
☐ DG Skill sets
☐ IT Department
☐ Other: □□□□
Executive participation in a DG program should be *

☐ From the start

☐ Not needed

☐ Partial, but not mandatory from the start

Is Data Management part of your *

☐ Project Planning (SDLC, Agile etc)

☐ Operational Support process

☐ Change Management process

Does the organization have *

☐ An active DG Steering Committee

☐ Reoccurring Data Quality / Management meetings

☐ Incentive or performance compensation for DG

☐ None of the above

In the next 24 months, will DG's importance *

☐ Increase

☐ Decrease

☐ Remain the same
Appendix B Table of Abbreviations

The following table contains a list of abbreviations used throughout this document.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BG</td>
<td>Business Glossary</td>
</tr>
<tr>
<td>AUM</td>
<td>Assets Under Management</td>
</tr>
<tr>
<td>BAU</td>
<td>Business As Usual</td>
</tr>
<tr>
<td>BG</td>
<td>Business Glossary</td>
</tr>
<tr>
<td>CDO</td>
<td>Collateral Debt Obligation</td>
</tr>
<tr>
<td>CDS</td>
<td>Credit Default Swap</td>
</tr>
<tr>
<td>CLO</td>
<td>Collateral Loan Obligation</td>
</tr>
<tr>
<td>COE</td>
<td>Centers of Excellence</td>
</tr>
<tr>
<td>DG</td>
<td>Data Governance</td>
</tr>
<tr>
<td>DMF</td>
<td>Data Management Framework</td>
</tr>
<tr>
<td>DQ</td>
<td>Data Quality</td>
</tr>
<tr>
<td>EDM</td>
<td>Enterprise Data Mart</td>
</tr>
<tr>
<td>EDM</td>
<td>Enterprise Data Management</td>
</tr>
<tr>
<td>EIM</td>
<td>Enterprise Information Management</td>
</tr>
<tr>
<td>ESB</td>
<td>Enterprise Service Bus</td>
</tr>
<tr>
<td>FpML</td>
<td>Financial Products Markup Language</td>
</tr>
<tr>
<td>GCDO</td>
<td>Global Chief Data Officer</td>
</tr>
<tr>
<td>GCTO</td>
<td>Global Chief Technology Officer</td>
</tr>
<tr>
<td>GISP</td>
<td>Global Investment Systems Platform</td>
</tr>
<tr>
<td>GRC</td>
<td>Governance, Risk and Compliance</td>
</tr>
<tr>
<td>GUI</td>
<td>Graphical User Interface</td>
</tr>
<tr>
<td>IT</td>
<td>Information Technology</td>
</tr>
<tr>
<td>MDM</td>
<td>Master Data Management</td>
</tr>
<tr>
<td>MDR</td>
<td>Metadata Repository</td>
</tr>
<tr>
<td>NAV</td>
<td>Net Asset Value</td>
</tr>
<tr>
<td>PUID</td>
<td>Pioneer Unique Identifier</td>
</tr>
<tr>
<td>SDLC</td>
<td>Software Development Life Cycle</td>
</tr>
<tr>
<td>SEC</td>
<td>Securities and Exchange Commission</td>
</tr>
</tbody>
</table>
References


http://www.infoworld.com/d/architecture/information-service-131


http://www.thetradenews.com/operations-technology/pre-post-trade-reporting/3585


managemen.com/news/1008514-1.html


Certification of Authorship of Thesis

Student Name: Patrick Egan
Student Telephone: 781-367-0623
Student Email: egan.patrick@gmail.com
Degree Program: MScSED
Advisor Name: Ernest Eugster, Ph.D.
Date of Submission: EFFECTING DATA QUALITY THROUGH DATA GOVERNANCE: A CASE STUDY IN THE FINANCIAL SERVICES INDUSTRY

Certification of Authorship: I hereby certify that I am the author of this document and that any assistance I received in its preparation is fully acknowledged and disclosed in the document. I have also cited all sources from which I obtained data, ideas, or words that are copied directly or paraphrased in the document. Sources are properly credited according to accepted standards for professional publications. I also certify that this paper was prepared by me for the purpose of partial fulfillment of requirements for the MCT 626 course.

Patrick Egan

08/26/2011

Student Signature Date

Revised 1/29/2010