

# Getting Started with Data Governance

Gregory S. Nelson

ThotWave Technologies, Chapel Hill, NC

## Abstract

While there has been tremendous progress in technologies related to data storage, high performance computing and advanced analytic techniques, organizations have only recently begun to comprehend the importance of parallel strategies that help manage the cacophony of concerns around access, quality, provenance, data sharing and use.

While data governance is not new, the drumbeat around this, along with master data management and data quality, is approaching a crescendo. Intensified by the increase in consumption of information, expectations about ubiquitous access, and highly dynamic visualizations, these factors are also circumscribed by security and regulatory constraints.

In this paper we provide a summary of what data governance is and its importance, which is, as you would expect. However, we go beyond the obvious and provide practical guidance on what it takes to build out a data governance capability appropriate to the scale, size and purpose of the organization and its culture. Moreover, we will discuss best practices in the form of requirements that highlight what we think is important to consider as you provide that tactical linkage between people, policies and processes to the actual data lifecycle. To that end, our focus will include the organization and its culture, people, processes, policies and technology. Further, our focus will include discussions of organizational models as well as the role of the data steward(s), and provide guidance on how to formalize data governance into a sustainable set of practices within your organization.

<b>INTRODUCTION .....</b>	<b>2</b>
RELATED CONCEPTS.....	2
<b>MEASURING DATA GOVERNANCE AND DATA MANAGEMENT .....</b>	<b>3</b>
THE PRINCIPLES OF DATA GOVERNANCE.....	5
MEASURING SUCCESS .....	6
COSTS .....	8
MEASURING RETURN ON INVESTMENT.....	8
<b>CAPABILITIES DEFINED.....</b>	<b>9</b>
<b>GETTING STARTED .....</b>	<b>11</b>
START WHERE YOU ARE.....	11
<b>SUMMARY .....</b>	<b>13</b>
ACKNOWLEDGEMENTS .....	14
REFERENCES.....	14
CONTACT INFORMATION .....	14

## Introduction

As an American, I am spoiled. Every day, I take for granted that when I go to the sink to wash my face or fill a glass, the water will be clean and healthy. On the occasion that the faucet sputters, is cloudy or simply does not flow, I am outraged. Behind the scenes, there are hundreds, if not thousands of people in support of that clean, potable, flowing water. In fact, even in my little town in rural North Carolina, I receive an annual report card that summarizes the chemical makeup of my drinking water as compared with the State and Federal guidelines. While I seldom read more than the masthead, I am comforted to know that it is someone's job to make sure that I know the results. Similarly, I enjoy the fruits of their labor.

For centuries, civilizations have continued to perfect the process of bringing safe drinking water to people and the used, "gray" water, back out of our homes and offices. For a fascinating read, I encourage you to take a look at *Flushed: How the Plumber Saved Civilization* (Carter 2006). While we've had a few centuries improve on the process of bringing water to our homes, we are still in our relative infancy when it comes to handling "clean, potable, consumable" data and ensuring its quality.

As data become more ubiquitous in our lives (think of the constant stream of information flowing into our lives through mobile devices), our expectations of data and how we use that for decision-making places increased demands on the "plumbing" within our organization. Our Data 'lakes' and 'oceans' require governance to ensure its variety, volume, velocity, and veracity are appropriately directed.

### ***Related Concepts***

In a previous paper (Nelson 2014), we outlined our concepts of data management and data governance, which often get confused. A number of important concepts are worth defining again:

- **Data governance** refers to the organizational focus that helps to ensure confidence around data and how it is defined in the context of the business – its use, interpretation, value and quality – is managed.
- **Data management** is the global set of practices that guide how data strategies are designed, executed and governed within an organization. Data management is the

guiding principles, architectures, policies, practices and procedures for managing data within enterprise.

- **Data integration** means bringing data from two or more sources together into a single view for analysis and reporting. The tactics of data integration include extract, transform and load (ETL) tools and techniques.
- **Data Quality** is the process of ensuring that data "are fit for their intended uses in operations, decision making and planning." (Juran and Godfrey 1999). As an overall framework, data quality comprises much more than software — it also includes people (roles, responsibilities, organizational structures), processes (monitoring, measuring, reporting and remediating data quality issues) and, of course, technology.
- **Master Data Management** is a discipline in which an organization comes together to ensure the "uniformity, accuracy, stewardship, semantic consistency and accountability of an enterprise's official, shared master data assets" (White 2009).

## Measuring Data Governance and Data Management

One technique that is often used is evaluating the value of data governance is to look at the missed opportunities within an organization. In our work with customers in healthcare, we often see the appetite for analytics outpace the organization's ability to support basic data needs. Here are a few examples of goals as well as the type of assets they require:

- **Center for Medicaid & Medicare Services (CMS) Meaningful Use Program**
  - *Goal:* Realize CMS financial incentives by achieving a number of capabilities in electronic health record (EHR) use. While one may initially think that certified EHR technology by definition should solve the reporting issues, there are a number of business directives that must be in place in order to receive credit
  - *Required Governance and Management Assets:*
    - Master patient index that uniquely identifies all patients of a hospital or eligible provider.
    - Master provider index that uniquely identifies all providers across an organization, even if they traverse care sites.
    - Appropriate and consistent usage of standards to be able to meet interoperability requirements

- **Geospatial Analytics for Population Health**
  - *Goal:* Augment the electronic health record or other patient data repositories with geospatially-derived overlay data such as the distance that a patient has to travel to see their provider; how close they are to a grocery store to access fresh food; and socio-economic information related to wealth, crime, and education.
  - *Required Governance and Management Assets:*
    - Business definitions and collection processes for different address types of patients (e.g. residential, mailing, seasonal) and care locations.
    - Process to define best guess of a patient's current address as well as archive address history.
    - Process to verify addresses in real-time against USPS or country-specific address sources.
    - Definition of geocoding precision acceptable for each analytic activity (e.g. tax parcel centroid, rooftop, street interpolation, zip+4)
    - Definition of processes to geocode new information as it is received
- **Management of Chronically Ill Populations**
  - *Goal:* Monitor and manage health outcomes and healthcare services consumption of those individuals with common chronic illnesses (e.g. type II diabetes, hypertension, or coronary artery disease).
  - *Required Governance and Management Assets:*
    - Structured electronic health record data for all chronic disease cohorts.
    - Business definitions that describe each chronic disease group in terms of what can be queried electronically. ICD and CPT billing codes are rarely sufficient.
    - A scheduled, automated process by which chronic disease cohorts are updated to include those newly diagnosed or those new to the health system.
    - A process by which to review definitions and update as medical knowledge grows over time.

- The ability to collect structured patient reported outcomes and family history data that can be categorized according to accepted data standards.
- Business definitions for agreed upon outcomes to monitor (e.g. readmission, length of stay, emergency department visit volume)

While this represents only a few examples, dozens of other analytic projects are often stifled by the fundamentals of good, quality data. Ask people in your own organization this simple question: *“What could you accomplish if you knew you had the right data and could trust that it was accurate?”* We asked this of just one health leader who responded immediately with nearly forty project ideas that could support patient care and executive decisions if only the right data was accessible. For example:

- Forecast emergency room visits
- Model equipment inventory and staffing needs using RFID tagging
- Notify study coordinators of potential recruits in real time
- Evaluate the impact of different health system initiatives on provider productivity
- Fine tune readmission rate models

### ***The Principles of Data Governance***

Perhaps you’ve never wondered about how we got clean, drinking water. The measure of success for clean water is fairly obvious: we don’t get sick. Similarly, in the world of the modern enterprise, imagine making decisions at a strategic or tactic level without the first knowing whether the data that you have put your trust in is worthy of our allegiance. Just as I assume that the process of delivering my drinking water has a system of checks and balances, planned and thoughtful oversight and strategies for risk mitigation, should we not assure ourselves that the same is true for data – the basis of all fact based, information driven organizations?

As a pragmatist, I find the work of Robert Seiner refreshing. In this book, *“Non Invasive Data Governance”* (Seiner 2014) talks about the four principles of data governance. These include:

- Data must be recognized as a valued and strategic enterprise asset
- Data must have clearly defined accountability
- Data must be managed to follow internal and external rules and regulations
- Data quality must be defined and managed consistency across the data lifecycle

In my opinion, clarifying the principles of data governance helps shed some much-needed light on how we measure its success. The four guiding principles of data governance above help provide the promise that if we do those things – if we treat data appropriately, if we have people accountable for its care, if we follow repeatable processes and we are consistent in those efforts – we will realize a benefit. For me, that benefit is confidence – an assurance that the foundation of all fact-based decisions is sound.

### ***Measuring Success***

The measure of success of any program is, by definition, in the eyes of the beholder. If an Information Technology (IT) department conceives, funds and manages the program, then the measures will likely be technical in nature. On the other hand, if functional areas (i.e., “the business”) oversee the program, then the measures will reflect areas that the business deems essential. Below, we have summarized some of the key benefits that may realize in a data governance program. We summarize this in the context of some key categories of governance.

**Table 1. Technical and Functional Measures of Success**

Category	Technical	Business
<b>Metadata</b>	<ul style="list-style-type: none"> <li>• Data flow metrics (# documented, monitored)</li> <li>• Data dictionary (# terms, # search term lookups)</li> <li>• Orphaned data assets (unused reports, data elements)</li> </ul>	<ul style="list-style-type: none"> <li>• Response time in determining whether a question can be answered</li> <li>• The percentage of data elements where the following can be done on demand: a) visualize the lineage of a data item and b) assess the downstream impact on other data elements if something changes</li> <li>• Number of data elements with an agreed upon business definition</li> </ul>
<b>Content Management</b>	<ul style="list-style-type: none"> <li>• Percentage of content that has been digitized</li> <li>• Percentage of unstructured metadata populated in the metadata repository</li> </ul>	<ul style="list-style-type: none"> <li>• Number and type decisions that are made (reinforced) with text analytics</li> <li>• Latency (in hours) to make a decision with contextual information</li> </ul>
<b>Archiving</b>	<ul style="list-style-type: none"> <li>• Total storage (in GB)</li> <li>• Total Cost of storage</li> <li>• Application response time</li> </ul>	<ul style="list-style-type: none"> <li>• A measure of our ability to reproduce historical reports and the context under which that was true</li> </ul>
<b>Security and Privacy</b>	<ul style="list-style-type: none"> <li>• Ratio of breeches by the total number of cyber attacks</li> <li>• Total number of records lost to breeches, negligence</li> </ul>	<ul style="list-style-type: none"> <li>• Reputational risk given data loss</li> <li>• Turnaround time for securely sharing data with researchers and other partners</li> </ul>
<b>Value Creation</b>	<ul style="list-style-type: none"> <li>• # of reports</li> <li>• # of data sources integrated</li> <li>• # of business elements defined</li> <li>• Ratio of new report requests by those completed</li> </ul>	<ul style="list-style-type: none"> <li>• Number and type of decisions that could be made given the integrated, high quality data (accuracy, validity, completeness, integrity)</li> <li>• How quickly one can answer a new or novel question (timeliness)</li> <li>• The extent to which analytic results can be reproduced across people and time and the consistency/ reliability of the interpretations</li> <li>• The velocity by which new talent can understand and utilize the enterprise's information assets</li> </ul>

## **Costs**

One of the reasons that organizations do not embark upon strategies of data governance is that it is perceived of as “too big and too costly”. Admittedly, it does seem daunting as it fundamentally means getting people to talk to one another (“communication is hard”), dealing with ownership issues (“she who owns the data, rules!”), and dealing with poor and inefficient processes (“pointing fingers”).

But as Bob Seiner points out in “Non Invasive Data Governance” (Seiner 2014) there are really five core messages that can and should be shared with senior leadership. These include (Note: I have taken liberties with my paraphrasing):

- A realization that most organizations already are already governing data (just informally.)
- You don’t have to blow up current organizational charts to do this, just provide structure and support to aid people to do this better.
- An acknowledgement that you can do things better
- This doesn’t need to cost a lot of money. Incremental improvements can be made by improving processes and supporting people in the job they are already doing.
- Use methodologies that fit the culture and needs of an organization. This begins with a shared vision and shared accountability about how data is to be governed.

## **Measuring Return on Investment**

While there are tools and technologies that can help with data governance, I am a firm believer in “small wins, big successes.” What I mean by that is that part of the tactics of any data governance program should include the measurement of progress. To do so means understanding how others measure your success. Some of these are easier than others – such as how long it really does take someone to become proficient as an analyst in your environment, whereas others will take some grey matter to figure out. (Nelson, Matthews et al. 2006)

Here are some suggestions for things that might go into calculating the return on investment for a data governance program:

- **Cost avoidance** – reduction in duplication of technology work/ activities; creation of assets that are unimportant/ unused; duplication of compliance related activities/



controls, elimination of duplicate systems/ technology (infrastructure, applications, data feeds).

- **Wasted Time/ Improved Time to Decision** – two sides of the same coin that can include waiting for data, spending time on non-value creating activities, churning through data quality issues or time spent finding and acquiring data.
- **Improved Communication/ Collaboration** – as data is recognized as an enterprise asset, others can benefit from the sharing of information, analytic results and knowledge about the interpretation and impact on action.
- **Cost of Poor Quality** – poor decisions that were made based on inadequate or inaccurate data; reductions made in erroneous decisions.
- **Improvement of Actionable Models** – improvement in analytic models and resulting impact/ decisions made (e.g., demand management forecasting and their impact on staffing models and patient volume, readmissions rate model targeting, utilization, etc.)

## Capabilities Defined

As technology becomes ubiquitous in our personal lives and across the modern enterprise, we are often quick to start researching technologies to help solve some of these problems. When we help customers with their data governance roadmap, we always start with identifying measures of success – call it ROI, benefits realization, or simply project justification. Part of that exercise is to craft a longer-term roadmap which necessarily involves defining the capabilities (Table 2) an organization needs to make governance successful. Questions that pop up include, *“how should we organize ourselves?”*, *“what technologies can be applied and in what order should be acquire them?”*, *“what talent / skills do we need and should we buy or build (contract versus grow)?”* We typically assess such organizational needs in order of ❶ Mindset; ❷ Skillset and ❸ Toolset.

Here we define those in the context of data governance:

1. **Mindset** (How the organization sees, perceives, and views data governance) – Across all organizational echelons, mindset describes the core set of beliefs, a consistent approach to prioritization, and realistic expectations of what can be accomplished. This also shapes how all stakeholders in the organization interpret the promise of data governance. Mindset also helps craft a set of guiding principles that inform how individuals will respond to situations.

2. **Skill Set** (Do, Act, Behave) – This involves defining the critical skills and behaviors required to deliver data governance. The level of expertise in a particular data governance activity determines the efficiency and effectiveness of an individual. This in turn translates into how the organization matures in their capability. Highly skilled individuals should be able to map the data generated from healthcare processes and evaluate process efficiency without the aid of technologies. This avoids the trap of being swept into suboptimal processes that do not fit the business, or worse yet, thinking that one is “done” despite the gaps in the technology solution.
3. **Toolset** (How you execute on the vision) – This is essentially the “how” of data governance. This means using a set of widely accepted methods, techniques, models, approaches, and frameworks that can create value. This can include collaboration technologies, shared statistical computing environments, data quality tools, business rules and metadata repositories and solutions for master data management.

**Table 2. Capability Requirements for Data Governance**

Organizational Capabilities	
<b>People and Culture</b> (Who)	<ul style="list-style-type: none"><li>• Consistent and broad awareness of the data governance vision, mission, purpose, and strategy</li><li>• Establishment of decision rights</li><li>• Clarity in data-related (roles and responsibilities)</li><li>• Definition of critical job competencies</li><li>• Training and development plans for change management</li></ul>
<b>Process</b> (What)	<ul style="list-style-type: none"><li>• Define, approve, communicate and align policies, standards, architectures, procedures and metrics</li><li>• Mechanisms for issue escalation/resolution</li><li>• Prioritization mechanism/ triage new requests with line of sight to strategy</li><li>• Establish, review, approve and monitor policies</li><li>• Processes for assessing risk and defining controls to manage risk</li><li>• Align data security architecture frameworks and initiatives</li></ul>
<b>Capabilities</b> (How)	<ul style="list-style-type: none"><li>• Data quality prevention, detection and resolution/ remediation</li><li>• Data stewardship including management, understanding and promoting the value of data assets</li><li>• Actively provision and manage master data (enterprise wide)</li><li>• Metadata management (create, manage, report, search)</li><li>• Track and enforce conformance to data policies, standards, architecture, and procedures</li><li>• Data privacy, protection, data sharing, and compliance with official data policies is actively enforced</li></ul>

## Getting Started

### *Start where you are*

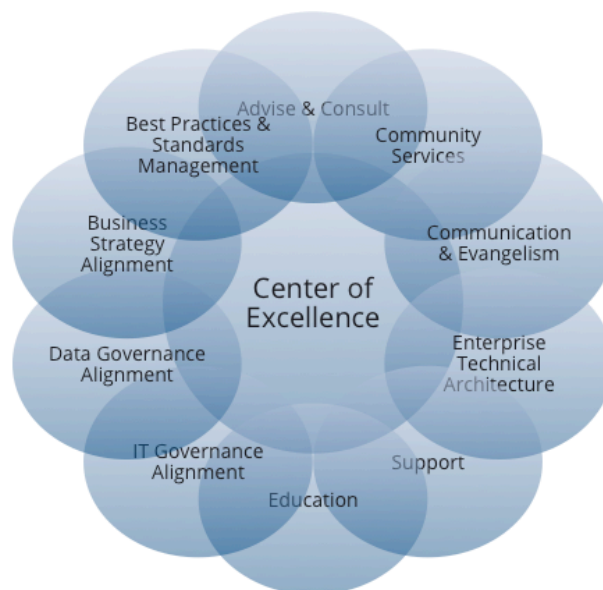
So you know what you want to do. You know that it is important and can see the benefits. So where do you start? As a consultant I am often embarrassed to say – *it depends*. But it really does depend on where you are. You can hire a consultant to perform an assessment or conduct an honest, internal evaluation of your own. Here, we would expect you to be able to articulate what best practices are, analyze your own strengths and weakness, and enumerate opportunities. The gaps in best practice and your current state can then be your roadmap.

For example, if your analytic teams struggle with getting complete, accurate data for their analytic projects, then I would probably start with an exercise which maps the information needs (data elements) categorized by subject area and data domain with the individuals/ departments/ systems that have been defined. The common data matrix, as it is referred to, does not have to be an enterprise wide exercise and take months to perform, but rather it is a tactical exercise that will get you started with one project, for one set of users, with a valid business priority.

With solid executive support, this can then grow to include other groups, departments, data sources and even other organizations. This matrix can evolve to provide evidence of the value of managing data through the complexity of a large organization.

As you go through this process, you will often find that you will have to address the capabilities of defining roles and responsibilities, policies, and engaging stakeholders from various parts of the organization in helping you solve these issues. There is always a tipping point at which this work becomes recognized and resourced beyond the efforts of a few. Having a strong champion and developing an ecosystem of raving fans throughout the business and technology communities that are affected by your work can have a tremendous influence on your work.

With that said, there are alternative approaches which involve creating organizational structures which consolidate already mature silos of competencies around data quality, MDM and data governance. In these cases, we often see people starting through the deployment of a data governance center of excellence (COE). There is no one right organizational model, but the diagram here highlights some of the focus areas for a center of excellence.



## Summary

In this paper, we have defined data governance from a healthcare perspective and described its benefits.

As we have noted, there has been tremendous progress in technologies related to data storage, high performance computing and advanced analytic techniques. However, attention to people, process and culture has largely been supplanted by the focus on technology and data.

We rely on being able to access, analyze, interpret and act on data in our organizations. As such, we should feel confident that the processes by which that data came to become part of the decision making process was sound and well governed. Furthermore, we want to rest assured that the data is well protected in light of recent news regarding data breaches, data loss and unauthorized access of data.

## ***Acknowledgements***

I would like to thank Monica Horvath and Kevin McNair for their “thot-ful” and insightful comments on this manuscript and for their continued support of the data governance work that we do in healthcare.

## ***References***

Carter, W. H. (2006). Flushed : how the plumber saved civilization. New York, Atria Books.

Juran, J. M. and A. B. Godfrey (1999). Juran's quality handbook. New York, McGraw Hill.

Nelson, G., M. Matthews and N. Howard (2006). Metrics Unleashed II: Measuring Productivity or Inspiring Artisans. Pharmaceutical Users Software Exchange. Dublin, Ireland.

Nelson, G. S. a. D., Lisa (2014). Modernizing Your Data Strategy: Understanding SAS® Solutions for Data Integration, Data Quality, Data Governance and Master Data Management SAS Global Forum, Washington, DC, Global SAS Users Group.

Seiner, R. (2014). Non-Invasive Data Governance: The path of least resistance and greater success, Technics Publications.

White, A. (2009). "Defining MDM – again." Gartner Blog Network  
<http://blogs.gartner.com/andrew.white/2009/07/01/defining-mdm-again/>.

## ***Contact information***

Your comments and questions are valued and encouraged. Contact the authors at:

Greg Nelson      [greg@thotwave.com](mailto:greg@thotwave.com)

ThotWave Technologies, LLC

1289 Fordham Boulevard #241

Chapel Hill, NC 27514 (800) 584 2819

<http://www.thotwave.com>

### **Greg Nelson, President and CEO, Thotwave Technologies, LLC.**

Greg is a global healthcare and Business Intelligence (B.I.) executive with over two decades of experience and leadership in the field. Greg is a prolific writer and speaker interested in healthcare analytics and the strategic use of information technology.

He received his BA in Psychology from the University of California at Santa Cruz and advanced his studies toward a PhD in Social Psychology and Quantitative Methods at the University of Georgia. Recently, Greg completed his Masters degree from Duke University in Clinical Informatics from the

Fuqua School of Business. His academic and professional interests include helping organizations mature their analytic capabilities. Founder, President, and CEO of ThotWave Technologies, a niche consultancy specializing in healthcare analytics, Greg is particularly interested in how lessons from across other industries can be applied to help solve the challenges in healthcare.

With certifications in Healthcare IT, Project Management, Six Sigma and Balanced Scorecard, Greg is also a prolific writer and has presented over 200 professional and academic papers in the United States and Europe. He won the Best Paper Award in 2013 at the Pharmaceutical SAS Users Group Conference and sits on the board of the SAS Global Users Group. In 2011, Greg was selected by SAS into their loyalty partner group. "This program acknowledges and supports individuals who are recognized experts in their fields and have a long-term relationship with SAS."

Married to wife Susan and living on a small "farmlet" in rural North Carolina, Greg is an avid woodworker, enjoys photography, rides a Harley-Davidson Motorcycle, and strives to be a lifelong learner.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

*thinking data*® is registered trademark of ThotWave Technologies, LLC.

Other brand and product names are trademarks of their respective companies.