Improving Enterprise Data Governance Through Ontology and Linked Data

By R.J. DeStefano

Submitted in partial fulfillment

of the requirements for the degree of

Doctor of Professional Studies

in Computing

at

Seidenberg School of Computer Science and Information Systems
Pace University

February 2016

Approval Page

We hereby certify that this dissertation, submitted by **Richard J. DeStefano** satisfies the dissertation requirements for the degree of Doctor of Professional Studies in Computing and has been approved.

Lixin Tao
Dr. Lixin Tao
Chairperson of Dissertation Committee

2/12/2016
Date

Dr. Charles Tappert
Dissertation Committee Member

2/12/2016
Date

Meikang Qiu
Dr. Meikang Qiu
Dissertation Committee Member

2/12/16
Date

Seidenberg School of Computer Science and Information Systems

Pace University

Abstract

Improving Enterprise Data Governance Through Ontology and Linked Data

By

R.J. DeStefano

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Professional Studies in Computing

February 2016

In the past decade, the role of data has increased exponentially from being the output of a process, to becoming a true corporate asset. As the business landscape becomes increasingly complex and the pace of change increasingly faster, companies need a clear awareness of their data assets, their movement, and how they relate to the organization in order to make informed decisions, reduce cost, and identify opportunity. The increased complexity of corporate technology has also created a high level of risk, as the data moving across a multitude of systems lends itself to a higher likelihood of impacting dependent processes and systems, should something go wrong or be changed. The result of this increased difficulty in managing corporate data assets is poor enterprise data quality, the impacts of which, range in the billions of dollars of waste and lost opportunity to businesses.

Tools and processes exist to help companies manage this phenomena, however often times, data projects are subject to high amounts of scrutiny as senior leadership struggles to identify return on investment. While there are many tools and methods to increase a companies' ability to govern data, this research stands by the fact that you can't govern that which you don't know. This lack of awareness of the corporate data landscape impacts the ability to govern data, which in turn impacts overall data quality within organizations.

This research seeks to propose a means for companies to better model the landscape of their data, processes, and organizational attributes through the use of linked data, via the Resource Description Framework (RDF) and ontology. The outcome of adopting such techniques is an increased level of data awareness within the organization, resulting in improved ability to govern corporate data assets. It does this by primarily addressing corporate leadership's low tolerance for taking on large scale data centric projects. The nature of linked data, with it's incremental and de-centralized approach to storing information, combined with a rich ecosystem of open source or low cost tools reduces the financial barriers to entry regarding these initiatives. Additionally, linked data's distributed nature and flexible structure help foster maximum participation throughout the enterprise to assist in capturing information regarding data assets. This increased participation aids in increasing the quality of the information captured by empowering more of the individuals who handle the data to contribute.

Ontology, in conjunction with linked data, provides an incredibly powerful means to model

the complex relationships between an organization, its people, processes, and technology assets. When combined with the graph based nature of RDF the model lends itself to presenting concepts such as data lineage to allow an organization to see the true reach of it's data. This research further proposes an ontology that is based on data governance standards, visualization examples and queries against data to simulate common data governance situations, as well as guidelines to assist in its implementation in a enterprise setting.

The result of adopting such techniques will allow for an enterprise to accurately reflect the data assets, stewardship information and integration points that are so necessary to institute effective data governance.

Acknowledgements

I am thankful to have been surrounded by people who provided me with encouragement, good counsel, and love throughout my journey and would like to extend my gratitude to the following individuals:

First and foremost, I must thank the “ladies” in my life. My wife Cynthia who, much to her chagrin, encouraged me to start this journey by saying that if I were to first wait for things to quiet down, I’d never do anything. Her patience, love, and occasional nudging have been the bedrock of this achievement for which I will always be thankful. To my daughter Giavanna who has taught me more about perseverance than anyone could. Achieving this goal is nothing compared to the challenges she has faced and overcome in her young life. To my little one, Gabriella, who was just an infant when I started this five years ago. She fills me with joy every day and the depth of her compassion always makes me realize what’s really important in life. And finally, to my mother, the late Jean DeStefano. While I’m sorry I could not finish this while she was here, I have felt her encouragement, prayers, and strength along the way.

I would especially like to thank my supervisor, Dr. Lixin Tao for his wise counsel, guidance, and patience with me through this process. His keen insights and ability to appreciate the complexities and challenges of technology in a professional setting has been a source of inspiration and the encouragement that I needed to see this through. To Dr. Fred Grossman, for having the forethought to craft such a program at Pace University, I am truly thankful for having the opportunity to pursue this milestone in my life.

As much of the genesis of this research has emerged from my professional endeavors, I’d like to acknowledge the support I received from individuals with whom I worked. To Mike DeMao, my prior manager at Goldman Sachs, in addition to learning an immense amount about data management from him, I learned how a good manager can still get their hands dirty, be curious, and have fun. To my current manager, Cindi Guthrie at Cablevision, who, possesses more confidence in me than I do. Her trust in my abilities has inspired me to always aim higher both professionally and personally. To my colleague, Laura Gorham, despite incessant pen clicking, her occasional edits to this paper, feedback, and sense of humor were instrumental in helping me see the light at the end of the tunnel.

And finally to my best friend from high school, Mike Kieran who always brings the levity at the right time and has been patiently waiting for me to finish so we can resume playing Star Craft.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The increasingly fast paced, decentralized, and competitive business environment of today has driven companies to focus on looking at corporate-generated-business-data ("data"), such as HR, financial, sales, and customer data, for example, as a crucial asset within an organization [1]. Companies have sought to improve efficiencies and gain opportunities by exploiting the data it holds and produces. Additionally, in an era of increased regulatory scrutiny, organizations need ensure compliance with regulations such as Sarbanes-Oxley, Basel II, CobiTm, and HIPAA, to name a few. It is therefore imperative for them to be mindful of what data they have, where it is going, who can see it, what data it, itself creates, and where it eventually resides. In order to mitigate the risks such as poor data quality, information leakage, and maximize the opportunities that are associated with the handling of corporate data, it is imperative that organizations establish a data governance framework [2]. Data governance is a set of processes that ensures that important data assets are formally managed throughout the enterprise. It ensures that data can be trusted and that people can be made accountable for any adverse event that happens because of low data quality. It entails putting people in charge of fixing and preventing issues with data so that the enterprise can become more efficient. Data governance also describes an evolutionary process for a company, altering the company's way of thinking and setting up the processes to handle information so that it may be utilized by the entire organization. It's about using technology when necessary in many forms to help aid the process. When companies desire, or are required, to gain control of their data, they empower their people, set up processes and get help from technology to do it [3].

The need to increase data awareness is a key component in the foundation of a data governance approach, as an organization cannot govern what it lacks a keen understanding of. In order to understand and govern the corporate asset of data, an organization must have an awareness into the information that describes the creation, format, usage, movement, and stewardship. Having concise and transparent view to the state of enterprise data will increase the effectiveness of data governance initiatives and in turn increase the awareness and quality of the business data itself. Having a view of corporate data architecture and a means to maintain it are in fact a core segment to the DMBOK [4].

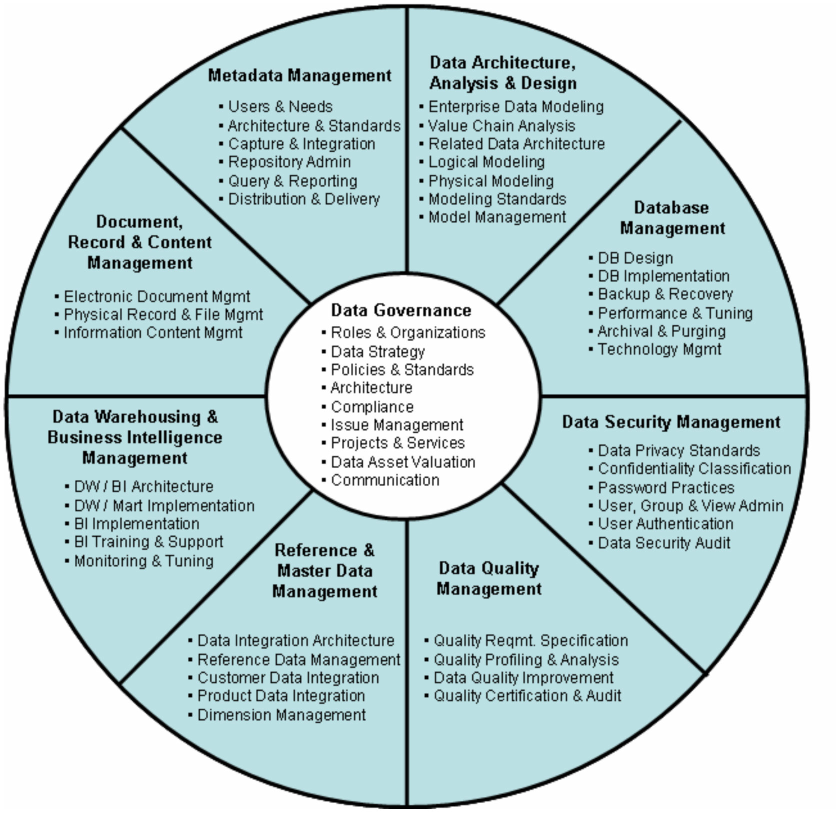


Figure 1 Data Management Body of Knowledge [4]

Unfortunately however, many large-scale organizations do not have a detailed, dynamic view of the data they have moving throughout their organizations and how it relates to the various functional groups and processes throughout the enterprise. The recent economic crisis in 2008 illustrated how ineffective knowledge of data hampered the ability to view, share the information, while making it more difficult for regulatory authorities to detect vulnerabilities in the global financial markets [5]. One of the key tenants to collaborative governance structure and sharing strategies proposed by, Pardo, Gil-Garcia, and Burke [6] is that there must be a knowledge of the environment to effectively disseminate data. For example, as large, multi-national corporate environments become more complex and data becomes easier to create at departmental levels, we start to see many new copies of data being created in order to side-step finding the true origin of the particular information. As

Pardo, et al state, the result is numerous “gold copies” of data which can be misleading when referred to by other groups within the organization [7].

Many large organizations do not have an awareness as to which business data are being created, where it is being used and duplicated throughout the organization, which are the appropriate stewards and stakeholders, where they reside in the organization, or what the impact of change is throughout the data supply chain. This lack of vision into the use of data relative to corporate structure and ongoing programs and portfolios contributes to overall data inconsistency, increased efforts to integrate disparate data, increased costs, and risk, while obscuring opportunities that may exist to streamline operations or increase revenue generation. All of which impede effective data governance.

Current solutions for increasing data awareness, such as master data management have taken the forms of commercial applications from IBM, Oracle, Informatica and many other niche companies [8, 9]. These approaches, while rich in functionality and industrial strength result in large-scale investments to the organizations in both time and money.

Additionally, frameworks, most notably the Open Provenance Model [10], with roots in the art and antiquities subject areas seek to create an implementation agnostic framework for mapping the provenance of any intellectual artifact or process. While this too is very rich and can model data in any form, certain subsets of the vocabulary may only be pertinent to organizations who are concerned with mapping meta-data contained in RDBMS throughout their organization. Additionally, other aspects of the vocabulary that are specific to organizational structure which are central to data stewardship are necessary.

It is the goal of this research to propose a cost effective, streamlined, and low impact approach to improving awareness of corporate data assets within large organizations by applying the principles of ontology, linked data, graph databases, and crowdsourcing in order to improve corporate data governance, thereby increasing the quality of business data and resulting in reduced risk and maximize opportunities. The scope of this research is focused on the data landscape within large, complex organizations.

1.2 Data Governance and Data Management

As stated in 1.1, data governance is a discipline, which focuses on the people, tools, and processes that are involved in the creation, management, and stewardship of enterprise data. It further refers to the parties "that hold decision rights and is held accountable for an organization's decision-making about its data assets" [2]. While IT Governance as a whole has mature frameworks, such as those proposed by Weil and Ross [11], data governance is still more fluid in definition [2]. Some of this ambiguity points to the fact that data governance is a practice that spans organizations, in which the business owns corporate data assets, as well as the policies that surround them. IT on the other hand is focused on data management, and the execution of data governance policies and decisions [1]. This partnership makes it imperative that all parties are aware of the corporate data landscape in a clear and concise manner. In the past when IT organizations and infrastructure were more centralized, keeping track of data ownership, movement, and consumption was a more straightforward task [12]. Data registries, data dictionaries, and data feed inventories were common tools employed to keep track of data movement. Additionally, reporting was a function that was more closely aligned to the IT department, so the data supply chain could be kept in check as IT was involved in the inevitable consumption of data by the business. This situation as with the overall corporate landscape however has evolved considerably into a more decentralized approach. First, organizations themselves have grown more diverse and complex [13]. Large global companies have many legal entities, customers, and suppliers and need to comply with local laws and maintain many source systems for similar data, additionally due to increased regulation, reporting has become much more important and often requires the consolidation of data from many diverse functional areas [14]. Next, as businesses become more decentralized, the creation of data has followed suit, resulting in groups no longer pulling data from a central source, rather many sources [7]. This has been further exacerbated by the advent of cheaper virtual hardware and the proliferation of desktop tools and techniques, such as MS Access and MS Excel, in which a team or small department can have their own infrastructure in place begin creating their own data to suit not only their purposes but can in turn supply a down stream process [15]. And finally, because of more sophisticated business intelligence tools, the business users now drive the reporting function. With the help of semantic "objects" which map to the underlying data, business users are able to extract and query data, thereby creating their own knowledge and dispensing it as they see fit. This evolution, while empowering the business to be more responsive to changing conditions and management requests, has made managing the flow of information more difficult, as uncoordinated, or independent approaches at data creation and definition by various groups in an organization results in data inconsistency, and ownership conflicts, thereby reducing the credibility of the data itself.

Closely intertwined with data governance, is the practice of data management ("DM"). A widely accepted definition of DM is "understanding the current and future needs of an enterprise and making that data effective and efficient in supporting business activities" [16]. Data management effectively supports and executes the strategy and policies put in place via data governance. Once again, as with data governance, we see there is a need for a "global view of corporate data" in order to effectively manage it [17]. Additionally, that view of corporate data must intersect with organizational and stewardship dimensions within the enterprise[18].

To reiterate, one needs to have a clear view of the data landscape in order to properly govern it *and* manage it.

1.3 Enterprise Meta-Data

While data management and governance is centered around appropriately managing and controlling the corporate asset of data in accordance with the strategic goals of the organization. One of it's primary inputs are the meta-data which describes the very asset that will be governed. Meta-data is data about data [2] and is crucial that it be available, accurate, and timely in order to appropriately make decisions on managing corporate data assets and enterprise knowledge. A meta-data repository is a pre-requisite in defining enterprise data architectures [19].

The following are key categories of metadata [20]:

- Entity / Attribute : At a very low level, this meta data describes the physical entities and attributes that make up data in business systems. This data can include descriptions of tables and columns and data type information in RDBM systems or files within organizations. Additionally, name, description, and data type of particular data elements, are very important when integrating with other systems or when developing applications on based on this data.

- Master Data / Transactional Data: This information indicates whether corporate data is master data or not. Enterprise master data is data that is central across the organization and used as a critical input into transactions or other business processes [21]. They are generally thought of as the "nouns" within the business such as account numbers, departments, customer and employee ID's. These data element are critical as they are generally used across multiple functions within the enterprise. It is therefore that their origin and propagation throughout the enterprise are well known. Transactional data is generally the output of systems or processes (though master data can be as well) which take master data as inputs and apply an operation on that data, such as selling products, or hiring employees, or executing trades. These data are generally what is aggregated and analyzed and reported on when monitoring the operations of a business.
- Provenance: Describes the creation, movement, transformation, and use of a data artifact. This meta-data is concerned with the ownership and lineage of data. This information is very important when identifying the flow of data within an organization or when trying to identify the transformations a piece of data can go through throughout its lifespan. The need to know the lineage of specific pieces of data has grown considerably due to increased external regulations, such as Sarbanes-Oxley, Basel I, Basel II, and HIPAA. Additionally, various countries have their own regulations as to where specific data may reside, and with whom.
- Stewardship: Meta-data which describes the parties within an organization and the role that they play regarding a specific piece of data. This information is increasingly important, as companies grow larger and more complex and decentralized. As

various smaller groups create, distribute, and consume data, the need to identify the parties involved in the creation and retention of data can drive many business decisions. Additionally, many organizations are seeing the benefit of combining data from multiple sources and functional areas across the organization. This horizontal aggregation involves many different parties and requires the appropriate stewardship to be put in place to effectively manage it.

1.4 Problems Faced Due to Poor Enterprise Data Awareness

1.4.1 Primary Problem: Poor Data Quality Through Inadequate Enterprise Data Governance and Management

High quality data are defined by the fact that they are “fit for their intended uses in operations, decision making, and planning”[22]. Poor data quality in organizations can take several forms, all of which compromise the confidence in which decisions or effective operations can take place based on the data in question. The primary forms are inconsistent data across an organization, data that is not accurate or incorrect, data that is latent and not timely, and data who’s source is unknown. The summation of these issues puts the organization at risk, impairs decision-making, missed opportunities, and results in re-work and higher costs [23]. Recent technological developments have enabled organizations to collect and store more information than has ever been possible. With the increased volume however comes increased complexity. This increase in both volume and complexity has given rise to challenges in managing it and increased risk of poor quality data [24].Well-defined data governance policies, along with robust data management, supported by a clear awareness of corporate data can help alleviate the impact of these situations.

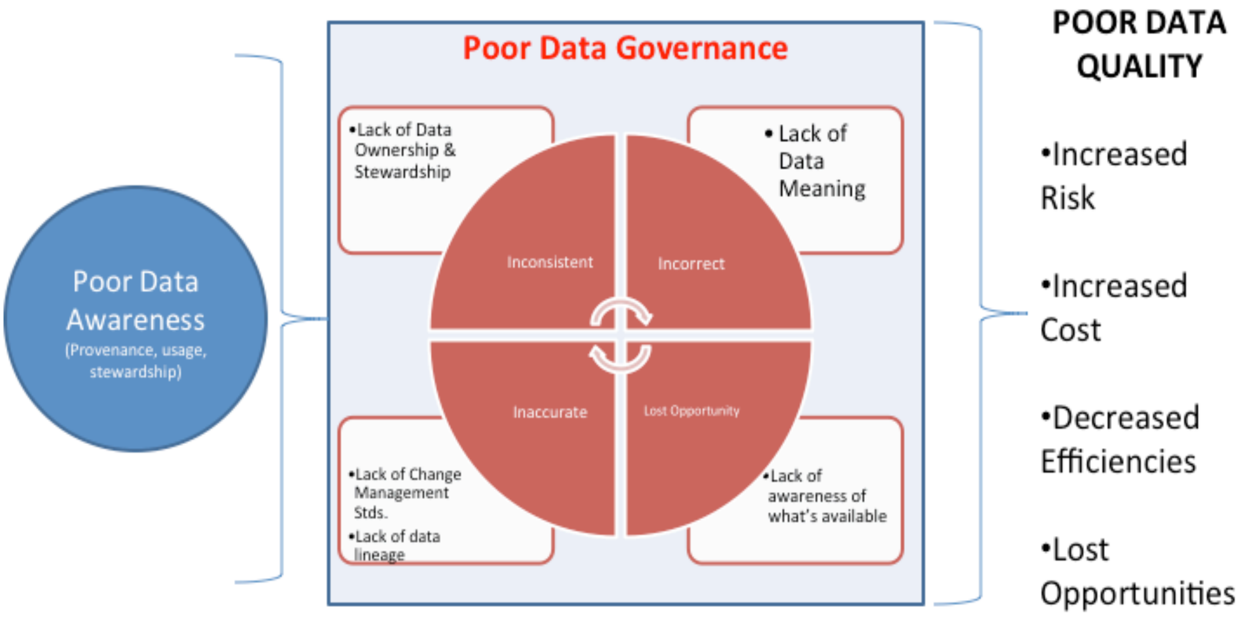


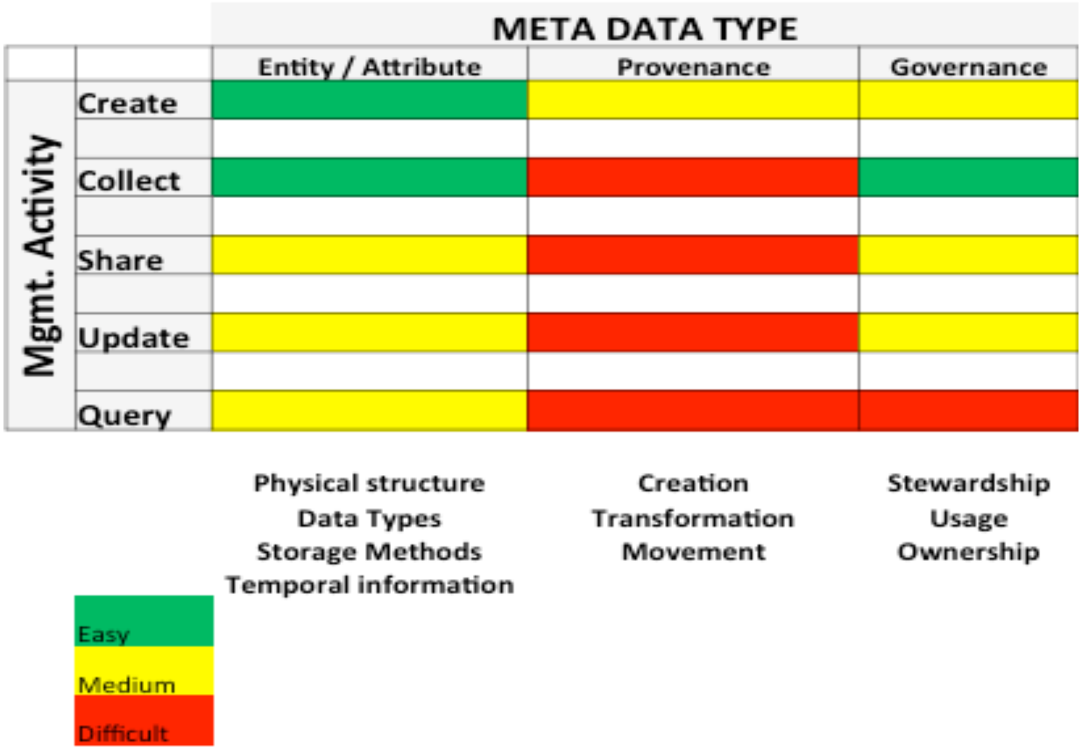
Figure 2 The Result of Poor Data Awareness

1.4.2 Secondary Problem: Implementing Enterprise Data Management Solutions.

In addition to the above challenges presented by poor data quality, implementing tools to increase data awareness and assist in enterprise data management has presented a further challenge. This is due to the fact that the actual prioritization of the required efforts are difficult to obtain because the benefits can be hard to quantify, thereby making a business case difficult [25]. Continued under investment fosters a siloed data landscape, thus leading to he business users generally who are not aware of the interdependencies of organizational data, except for when things go wrong. Adding to this, is the fact that implementing a rich solution requires participation across all stakeholders throughout the lineage of the data (both horizontal and vertical). This high level of organizational commitment to get started tends to result in a high barrier to entry to enterprise data management initiatives. Executive sponsorship and funding are key requirements for success for these enterprise wide data centric programs, as seen at IBM [19].

1.4.3 Fundamental Attributes of Data Awareness Challenges

The challenges that surround data awareness parallel the issues faced with implementing data management initiatives and are primarily driven by two dimensions: the type of data in question as well as the activities that surround managing that data. The activities involved in metadata management include: creating, collecting, sharing, updating, and querying. Different types of metadata within organizations pose different challenges; however as the dimensions are intertwined, effective metadata management requires addressing them all:



		META DATA TYPE		
		Entity / Attribute	Provenance	Governance
Mgmt. Activity	Create	🟩	🟨	🟨
	Collect	🟩	🟥	🟩
	Share	🟨	🟥	🟨
	Update	🟨	🟥	🟨
	Query	🟨	🟥	🟥
		Physical structure Data Types Storage Methods Temporal information	Creation Transformation Movement	Stewardship Usage Ownership

🟩 Easy
🟨 Medium
🟥 Difficult

Figure 3 Efforts Involved in Collecting Enterprise Meta-Data

Current approaches to enterprise data management have relied upon the said techniques of inventories and registries, but also many proprietary and open source tools. Most approaches however also involve a core set of individuals that drive the project and solicit information from the various data stakeholders and consolidate it, making assumptions about the movement and nature of the data. In a recent literature review on Data Warehouse Governance in healthcare settings, the primary approaches called for establishing a top down/centralized approach to data governance and management [26]. While these approaches can be successful as they afford more control over the project, it misses out on the ability to harness the contribution of many people throughout the organization through crowd sourcing input as to the stewardship, definition, and movement of data. This broad based participation throughout the enterprise is crucial to identifying and mapping the flow and usage of data throughout then enterprise, as different areas of the organization view, consume, and create data in different ways [27]. Furthermore, where stewardship is lacking, increased participation can foster an increased sense of ownership in the process and data.

Additionally other efforts regarding enterprise data management have focused on the "vertical data supply chain" whereas, more effort is placed on the semantic, or domain modeling aspects, with the intent on distributing the data via a service model. While this approach focuses more on the "domain-specific-metadata" [2] and information abstraction, rather than the data itself, it also tends to result in many governance issues as to what the true business definition or interpretation of the data should be, who owns it, who is consuming it, where is it being persisted, etc.. all of which need to be resolved before abstraction and distribution can be tackled [28]. Furthermore, it is reliant on identifying data sources, from which to abstract data from, which can be in question if many of the metadata challenges in this research are not met.

1.5 Scenarios Depicting the Need for Increased Enterprise Data Awareness

The following use cases illustrate situations that contribute to poor overall data quality. Observe the following use case for the Acme Corporation*:* Acme Corporation is a large global manufacturing company with over 100,000 employees, thousands of departments and corporate systems, and offices and plants in several countries. The corporation has many different functional areas, most of which both consume and produce data. As market pressures increase, they are continuously looking to analyze data across many business units in order to find opportunities or cut costs. This need for such cross functional data has given rise to many projects and efforts focused at identifying which teams have access to what data, and what they produce, enrich, modify, and consume. Some of the key data producing/consuming divisions are: manufacturing; human resources; finance; sales; product systems; and marketing. In addition, external entities, such as auditors in the multiple countries where Acme operates would need to see their data, as well as the controls around it. Having a robust sense of enterprise data awareness will assist the corporation in taking action and making informed decisions regarding its data.

- Example 1 – Impact Analysis, Proactive: The core HR system uses a 2 byte integer for the employee id field. They do not re use employee ID's so they can keep the history of terminated employees. They are approaching the maximum size and will soon run out of space. HR sends the definitive employee list to the finance department daily so they have a roster of all active employees daily for expense purposes, who then distributes it to other groups. HR also sends employee id data to sales so they have a roster of sales people across the regions weekly as well. There are also various applications and reporting systems that are accessing the HR database as well as sales and the general ledger system. HR needs to see the impact of increasing the length of the column or even the data type on receiving applications and databases. They also want to know who the stakeholders are that need to be involved in this activity.

- Example 2 – Impact Analysis, Reactive: Due to some internal error, the Product Information Management System (PIM) inadvertently had the active flag set to TRUE on the Product Master Table to a large group of discontinued products that were originally sold globally. While the Product team was able correct the error, it took them over a day to identify the issue. Since then, many downstream systems have received product data. The product team needs to know which processes and systems were affected and which people to contact. Management would like to know the impact from a regional point of view.

- Example 3 – Forensic Analysis: European regulators want to see all the confidential data about employees and financial information on certain customers that are being created in systems in the EU. These data can be created in local HR systems or CRM systems that support a particular region or office. They want to know where it is flowing, if it's encrypted, and if and when it gets destroyed.

- Example 4 – Identifying/Enforcing Business Definition Consistency: A small team is building a system that monitors customer satisfaction. Several groups such as finance and sales and marketing have data elements called "customer" however the Sales team's "customer" contains potential ones, whereas the finance teams have only "customers" who have made a purchase. They need to contact the owner of that data to confirm what is the true definition of "customer". Additionally, other systems such as Sales, use the phrase "sales" to refer to the amount of sales which took place in a certain time period, whereas in finance it refers to only those which have been realized.

- Example 5 – Managing IT Costs: Management needs to cut IT costs. They want to see which offices are consuming the largest amounts of data as well as identify redundant data stores, or identify areas to streamline data distribution. For example, manufacturing plants may not need to consume sales revenue data from other regions, or at a macro view, groups across the firm could be receiving Product data from many different sources because. Additionally, a particular software platform may be end of life or in need of an upgrade. The ability to see across the organization a consolidated view of the data environment can help Acme corp better plan and reduce IT expenses.

- Example 6 – Risk Mitigation: In order to reduce the liability of a potential data leak, a group in finance would like to identify all sensitive information about customers, who is using it and where it is being stored. They would like to purge redundant data stores that aren't necessary and have groups access a central repository.

- Example 7 – Opportunity Identification: As companies grow bigger and more complex, many groups are becoming data creators, and consumers. Consequently, this decentralized flow of information can provide important insight to senior management and show opportunities that may exist. For example, if a particular office within Acme Corporation is showing an increase in sales that is above average for that region, they may be utilizing more corporate data in a way other offices are unaware of. Looking at their data consumption can offer opportunities for senior management.

- Example 8 – Buy vs. Build Decisions: One of the most significant decisions a senior manager in information technology can make, when commissioned to implement a solution, is whether or not to purchase a third party tool and configure it to the organization's needs, or architect and implement a solution from scratch. Both approaches have their merits and an analysis of the alternatives is not within the scope of this research. Regardless of the approach taken however, most likely the solution will need to be integrated with legacy systems. Often, integration is a key driver in the decision making process. Having a view of all of the associated data, and the dependent systems, parties, and initiatives that are somewhat related to that data will make the decision and subsequent implementation more efficient.
- Example 9 – Process Impact Assessment: Often when there are corporate initiatives to streamline processes, or analyze them for audit purposes, having a clear understanding of the data and systems that are affected by that process are essential in describing potential impacts to the business should there be system or data issues. Conversely, when altering systems, it is important to know which processes than can be affected in the exercise so the appropriate mitigation measures can be taken. For example, the onboarding process in large organizations, where a new employee is brought into an organization is a process that has many distinct systems to bridge and much data passing back and forth. Being able to know the impact to systems when

contemplating process changes, and visa-versa will be highly beneficial in reducing risk to business operations.

1.6 Problem Statement

1.6.1 Research Problem

The essence of this research is to address the problem of poor data awareness in large organizations. Put simply, one cannot govern what one doesn't know. This is precisely why governments perform a census from time to time. When an enterprise lacks an awareness of it's data assets and how they relate to organization, it's ability to govern and administer those assets is compromised, leading to data quality issues across the organization. Organizations have at times attacked the problem through more compartmentalized attempts, such as within a department or a division. Often, these efforts correspond to data centric projects, such as data warehousing or business intelligence initiatives, however a comprehensive view of the corporation's data and how it relates to the organization is needed to provide an enterprise solution. This solution needs to provide an accurate view of the state of corporate data at any time. In order to do so, corporate meta-data must be collected and stored in a manner that is scalable, sustainable, and queryable. To address this, this research proposes an extensible framework in order to capture, store, and disseminate this information. This level of flexibility is critical in fostering adoption across the enterprise. Ideally, such a solution would be used and embraced by all teams, both functional and technical, however it is realized that some efforts need to start in a more contained manner to build momentum. The proposed framework fosters adoption by allowing users to participate at varying degrees, thus lowering a barrier to entry. Additionally, commercial data governance solutions such as those proposed by Informatica, Oracle, or Colibre tend to have high costs, complex implementations, or a ridged structure associated with them, providing more barriers to adoption.

The assumptions of this research are that the organizations in question are large and complex, whereas a complete view of data movement and stewardship is not possible without some form record keeping.

1.6.2 Solution Methodology

This research focuses on improving an organization's ability to govern and manage their data by addressing its “data awareness” through the use of linked data, semantic web techniques, and crowd sourcing, in order to leverage maximum enterprise participation via a low cost, incremental approach at modeling the data supply chain and its associated stewardship**.** The solution to this research problem is achieved through the construction of a framework, which will be tested against various use cases. The key components in the construction of this framework are as follows:

- Model: An ontology was constructed based on core data management concepts, that models the organization as well as key business processes and shows the manner in which data assets relate across the enterprise.
- Test Data: A suite of test data was created in order to validate use cases against.
- Extension: The ontology was also constructed in a way to illustrate the extensibility and ease of publishing from a de-centralized point of view.
- Extraction and Reporting: A series of use cases or business questions have been modeled in a querying language and applied to the framework.

- Provide an approach for industrialization: Here, we will define the approaches and architecture needed to make the framework feasible to implement, scalable for to parallel business change, and sustainable from a maintenance standpoint.

1.7 Overall Solution Architecture

1.7.1 Characteristics of Linked Open Data that Make it Suited for Enterprise Meta-Data Management.

The semantic web, or "web of linked data" refers to data that is published on the web is such a way that it is machine readable, it's meaning is explicitly defined, it's linked to other external data sets, and can in turn can be linked to from external data sets [29]. While HTML captures special data items through the use of tags to help convey additional information about particular concepts it fails to capture the relationships between multiple concepts. Linked data however enables a higher level of expressiveness in that it conveys not only attributes about the concepts themselves but how those concepts and attributes relate to other concepts and attributes. Relationships are the very fabric of linked data [30]. Whereas object oriented solutions are concerned with data and behavior confined to that object, relationships are secondary. In linked data, relationships themselves have their own inheritance and restriction rules and are independent of the objects themselves. Linked data is based on the Resource Description Framework (RDF) and Web Ontology Language (OWL) which are in turn are more expressive extensions of XML. Whereas XML structures data in a tree like format, RDF is in a graph structure that is focused on assertions that are comprised of triples where data elements are stored in a subject, predicate, object format. These statements can be linked with other triples and exposed as URI's so that other RDF graphs can reference them. Graph structures have been used to describe supply chain management which has a similar structure to the flows of data within an organization [31, 32]. Graph structures have been described also in terms of cataloging master data [33].

This expressiveness in a machine readable form coupled with the flexible nature of a graph database creates many powerful opportunities. First, applications no longer need to account for the lack of expressiveness of the data they deal with by containing it within its own logic. Rather, the data is self-describing and can convey a good deal about itself and the data it is linked with, thereby separating it completely from the presentation layer. Secondly, the use of HTTP as a standardized access method and RDF as a standardized data model simplifies data access. Third, because the linked data is open applications needn't be developed against a fixed set of data; rather they can discover new sources of data at run time.

This incremental approach of being able to link disparate sets of data is one of the main strengths of linked data for modeling data movement and governance within an organization. It allows for incremental growth and the ease of publishing models via referencing URIs. This low barrier to entry will foster enterprise participation through crowd-sourcing. As seen with dataspaces, linked data provides for a "pay as you go" [28] approach to modeling. This has benefits in several areas in the area of enterprise data management. First, it creates a low barrier to entry to such a project. As stated before data management projects tend to be large efforts where the ROI is difficult to quantify to the business. An economical approach that allows the organization to focus on specific areas at a time is generally more palatable for business and executive sponsorship. Second, it allows for the multitudes of people / groups to publish their own data and link to others. This allows for a more dispersed effort that is not only more likely to have corporate backing, but more timely and accurate information as the individuals who are handling the data will be participating directly in the modeling effort for their space. Third, the machine-readable nature of RDF allows for powerful engines to infer much more information that is stated in the initial assertions. These reasons, combined with the low cost aspects of HTTP access make semantic web techniques an appropriate candidate for modeling the movement of enterprise data. Additionally, once the corporate data landscape is modeled, queries can be executed on the graph to show the state of the data throughout the organization.

1.7.2 Solution Methodology: Addressing the Problem of Overall Poor Data Quality in Organizations.

As discussed, this research provides a framework for addressing overall data quality by addressing two of its primary contributing factors: poor data awareness and poor data governance. The methodology is to construct a framework that an enterprise can use to incrementally track corporate data assets, their flow, associated stewardship, as well as the relations to application portfolios. To do this, we will be employing an ontology to conceptualize aspects of data management that we want to capture. Additionally, the concepts of linked data provide a flexible foundation on which to collect and disseminate information and to build on and expand that ontology. That flexible foundation will also allow multiple individuals in the organization opportunities to contribute in a crowd-sourced fashion to ensure maximum participation and ownership of the initiative. The graph structure of the linked data foundation also provides an insightful and flexible way to store, query, and visualize the information. The transitive nature of graph databases allow for a robust way to query and extend data lineage. The framework will employ linked data and dataspace principles that will for an incremental approach at capturing enterprise meta-data in a graph structure that can be queried as well as expanded with the growth of the company.

Automated mechanisms will collect meta-data information from the various structured data throughout the firm and "seed" a graph structure based on an ontology that brings together the complex relationships between an organization, it's processes, and it's data that will discuss in detail in chapter 4. This can help serve as a jump-start for an organization with very little investment. In our case, while we embrace the flexible nature of linked-data and semantic web technologies, certain restrictions must be in place in an enterprise setting, such as a core, shared ontology, which provide for consistency across the organization, which would otherwise not be necessary in the semantic-web itself. The graph itself will be addressed in several ways in order to keep it current with the changing structure of the organization:

1) A "core" ontology has been defined that describe the organization, relationships, and key data governance concepts. Groups will be encouraged to build upon this core in a manner that will allow them flexibility while maintaining consistency.
2) The crawling mechanisms, or agents, can also serve as an auditing feature to detect any differences in physical (actual) database structures and the meta-data graph itself which represents those.
3) Manual intervention, IT teams can manually update the graph with additional information regarding the data structures and physical meta-data. Concepts such as stewardship and data governance policies will need to be updated manually, or via a custom interface.

4) In order to keep such a graph current and accurate, building it into the organization's software development lifecycle is imperative. As new applications are developed and existing ones are enhanced, structural changes, and data activity should be updated in the graph as part of a move to a production environment. This assumes appropriate testing and vetting of the modifications to the graph are correct.

1.8 Conclusion

1.8.1 Contribution

The contribution that this research will provide is an enterprise wide framework which will improve data awareness, data governance, and in turn corporate data quality while remaining practical and inclusive of many other individuals who would otherwise not take part in a highly centralized meta-data management approach. Teams can focus on their data; it's immediate sources, and immediate dependencies. The nature of linked data will allow for connections with other sources to in essence form a data chain. As with crowd-sourced initiatives, the framework will get richer and add more value as adoption increases, while incurring a minimal upfront investment.

1.8.2 Roadmap

The next chapter will provide a survey of the key topics in data governance and management, along with various technologies that have been used for meta-data management as well as ones which be employed throughout this research. Chapter 3 will summarize the solution methodology and the various components necessary and provide guidance on operationalizing the solution. Next, chapter 4 will describe implementation highlights across several different technologies. Chapter 5 will describe use cases, their results, and describe the set up and nature of the test data. Finally, chapter 6 concludes this paper and provides avenues for further research.

Chapter 2

Current State of Research

2.1 Data Quality

2.1.1 Background

The primary driver of this research is to improve corporate data quality by providing a framework, which lowers barriers in to manage the process in which data is created and distributed throughout large organizations. The importance of data quality is well established. Data quality is a key prerequisite in many key business operations and objectives across many industries. Examples of such are: global supply chain management [34], customer relationship management [35], strategic decision making and business intelligence [36], and, perhaps most importantly in the current business environment, compliance with regulations [37, 38].

As mentioned in chapter 1, data is deemed high quality if it is fit for its intended use. While there are different definitions for "fitness", it is commonly implied to span the following 4 categories [2]:

- Accuracy: Determines the correctness of the data as it was recorded, against its actual value, based on the context of its usage.

- Timeliness: The recorded data must be current for its intended use.

- Completeness: The data values are persisted and it is adequate of depth and breadth. For example, this could mean having all the history for a given account, or all of the data values for a particular product.

- Credibility: Pertains to the trustworthiness of the data source.

Unfortunately however the problem of data quality has not subsided, despite increased levels of time and funding. While there has been a significant amount of research into data quality methods and maturity models, a considerable amount has been performed by practitioners [39] and tends to be more rooted in professional experience and is anecdotal [32]. Two studies show how this problem is not getting any better. In 2002, a data quality study based on interviews with industry leaders, experts, and survey data from 647 respondents estimated the cost of data quality at USD $600 Billion per year [40]. Seven years later, a comprehensive study performed by Larry English quantified the problem on a global level at USD $1.2 trillion [41]. Within seven years, the cost of poor data quality had more than doubled. Those years have seen an explosion in the amount and complexity of data, corporate structure, and regulatory requirements. Clearly, methods to manage and govern that data have not kept pace with the evolving corporate environment.

Data quality issues themselves can span several dimensions from accuracy, timeliness, completeness and credibility [42]. Adding to some of the difficulties in managing data quality is that it is not an absolute concept as the quality of data is relative to the consumers of that data's needs. An example of this is seen where budgeting and decision-making applications may not need data in as timely of a manner or as precise a format as a line of business system, such as a trading or invoicing application. It is therefore essential to understand the organizational structure and processes behind this data as it moves throughout the organization. According to Watts et al., this point regarding the context of data use has been largely ignored by past data quality assessment models [24].

It is important to note however that data quality issues are further exacerbated in large multi national organizations. As stated by Huner et. al "Such companies possess a diversified portfolio of data storing and processing systems due to a history of mergers and acquisitions, deviant requirements of business units, and different regulations across countries" [39]. The data quality issues primarily arise when collecting data across business units and other organizational boundaries [43]. Where different groups not only use the data differently but perceive it differently as well. This point reinforces the limited target population for this research to large, complex organizations.

2.1.2 Impacts of Poor Data Quality

According to English, poor data quality itself can manifest itself in the following costs [23]:

- Process failure costs: These are associated with work functions creating and or transmitting "bad" data, causing another system or process, such as management reporting, billing, production, or trading to behave not according to their specifications, and in turn further propagating the transmission of bad data to other processes.

- Information Scrap and rework: these are costs that are associated in rectifying the issues from process failures as well as remediating the code that generated he said bad data.

- Opportunity Costs: This refers to the lost revenue from opportunities that were not pursued. A primary cause of this is that manpower and talent are dedicated to the above rework. Additionally, when bad data manifests itself into reporting and decision support systems, improper decision could be made, or the business environment may be incorrectly viewed, leading to faulty assumptions and strategy taken by the business.

Poor data quality also has the intangible effect of eroding trust and confidence in the data, possibly resulting in a reluctance on the behalf of the users to accept initiatives based on that data [44]. This can result in the proliferation of “spread marts” and “shadow technology” initiatives on the user side as they employ desktop tools and independently gathered data to generate knowledge [15].

Regarding studies in data quality, companies, industry experts and practitioners such as Gartner, Price Waterhouse Coopers, and The Data Warehousing Institute have been the primary contributors in the past. The following key outcomes have been summarized across several of those efforts [45]:

- 88 per cent of all data integration projects either fail completely or significantly over-run their budgets

- 75% of organizations have identified costs stemming from low quality data

- 33% of organizations have delayed or cancelled new technology initiatives because of poor data.

- According to Gartner, bad data are the primary contributor to CRM system failure.

- Less than 50% of companies claim to be very confident in the quality of their data.
- Business intelligence initiatives often fail to meet their objectives due to low quality data. This has the added effect of key business decisions being made on the outcome of such systems.
- Customer data typically degenerates at 2% per month.
- Organizations typically overestimate the quality of their data and underestimate the cost of errors.
- Vast amounts of time and money are spent on custom coding and traditional methods – usually firefighting to dampen an immediate crisis rather than dealing with the long-term problem.

Much of the problem we see is that data quality has been seen more of as a local problem [46]. Often times small groups or departments are able to deal with these issues by making manual correction, or just being aware of any anomalies. However, when cross organizational data sharing is required, as is the case with large ERP, Data Warehousing, or Business Intelligence initiatives, or an urgent regulatory data gathering exercise, the issues surface quite quickly. Unfortunately, as seen from the examples, the issues surface at times when there is high visibility on a project or when the data is need with the most urgency.

Having a global view of data movement may not be able to solve data issues, however a map of organizational data will provide a framework for annotating troubling combinations and identifying the people, organizations, and systems in the chain, in order to reduce the risk as well as accelerate any required remedy.

2.2 Data Governance and Enterprise Information Management

2.2.1 Background and Drivers

As we have seen how IT Governance parallels Data Governance [2], it is established that governing corporate data assets is not a technology solution, rather an ongoing behavior that organizations must adapt in order to manage the people, processes, policies, and tools around data in order to create a consistent view of organizational data [47]. Why do organizations need to take such steps? As stated earlier in this chapter, corporate data quality has become more and more important for companies in this age of an increasing data footprint. As data assets grow and companies become more diverse, and data collection methods become more ubiquitous, a more robust attitude to managing these assets is essential to increasing or maintaining its quality. Generally, as companies grow there is very little strategic thinking as to managing information. Often, as the organization is growing, it is easier to deal with these situations on a tactical basis and add “band-aids”. According to Sarsfield, “It’s usually cheaper and easier to fix the problem with glue and duct tape than it is to think about it strategically” [3]. As the organization matures however, the need for a concise strategy evolves as data siloes start to be created and timing and semantic anomalies start to occur. In most cases, as companies are smaller data is controlled by virtue of the same people acting on, consuming, and creating all the data needs within the organization. With growth, however, these activities become fractured across different teams, performing different functions within the organization, and possibly residing in different geographic locations. As systems that create and collect data are constructed, each leaves behind a legacy of users that are dependent on that data. While data integration for large scale ERP systems initiatives are generally accounted for, many smaller systems that are created which users will depend on are done so outside of a general IT, let alone data governance initiatives. Additionally, a lack of individuals that are skilled in the areas of data integrations within most organizations, exacerbates the problem [48]. Adding to the need for data governance practices, as companies grow through mergers and acquisitions, entire functions, teams, system architectures, and data flows must be integrated with the pre-existing environment. Having clear data governance policies will reduce ambiguity in determining the data, which needs to be integrated, as well as identifying the key parties involved.

As seen in chapter one, some of the key drivers behind establishing a robust data governance environment are internally driven, such as increasing data quality, more efficient operations, etc. There are however numerous external drivers to increased data governance practices. These are in the form of increased efficiencies in completing corporate actions, as discussed before with mergers and acquisitions. The other is driven by regulatory concerns [49]. Since the financial crisis, there has been increased scrutiny on the creation and storage, and securing of corporate data assets. In the wake of the Enron scandal, the Sarbanes-Oxley Act in 2002 was landmark in its data governance requirements placed on organizations via stringent internal controls on the creation, usage, and storage of data [50]. Additionally, BASEL II defines standard operating procedure in Europe which incorporates transparency in the data lifecycle [3]. It is also widely accepted that scrutiny on corporate data management practices will increase in parallel with increased regulation [51, 52].

Since 2004, IBM has established the Data Governance Council in order to address best practices and issues surrounding corporate data assets. The council has grown to over 50 major organizations across multiple industries [53]. While we will discuss the IBM Data Governance Council Data Governance Maturity Model, it was noted by NASCIO, which represents all of the CIO's of the state governments of the United States that in 2008, the IBM Data Governance Council made the following predictions about corporate data assets over the next four years [54]:

- Data governance will become a regulatory requirement.
- Information assets will be treated as an asset and included on the balance sheet.
- Risk calculations will become more pervasive and automated.
- The role of the CIO will include responsibility for data quality.
- Individual employees will be required to take responsibility for governance.

Aiken and Gorman's emphasis in 2013 on the need for and tasks of the Chief Data Officer in organizations has clearly affirmed these increasingly present regulatory drivers behind the need for data governance and enterprise information management [32].

2.2.2 Data Governance Maturity Frameworks

As with other processes within information technology managing data assets within an organization fits well with the Capability Maturity Model concept in which it is clearly stated that: "The benefits of better methods and tools cannot be realized in the maelstrom of an undisciplined, chaotic project" [55]. In the world of data management, we can replace the word project with "environment" as data management is a function, which pervades both technical and non-technical projects. While there are elements about data management which set it aside from software development, such as the management of it's persistence while being created, accessed, and "at rest", as well as identifying all of the parties across the organization that are involved in it's lifespan, the position of Paulk, et al. that the best tools and processes will be of little value if there is little capability, or appetite for that matter in embracing improvement, easily spans across both disciplines.

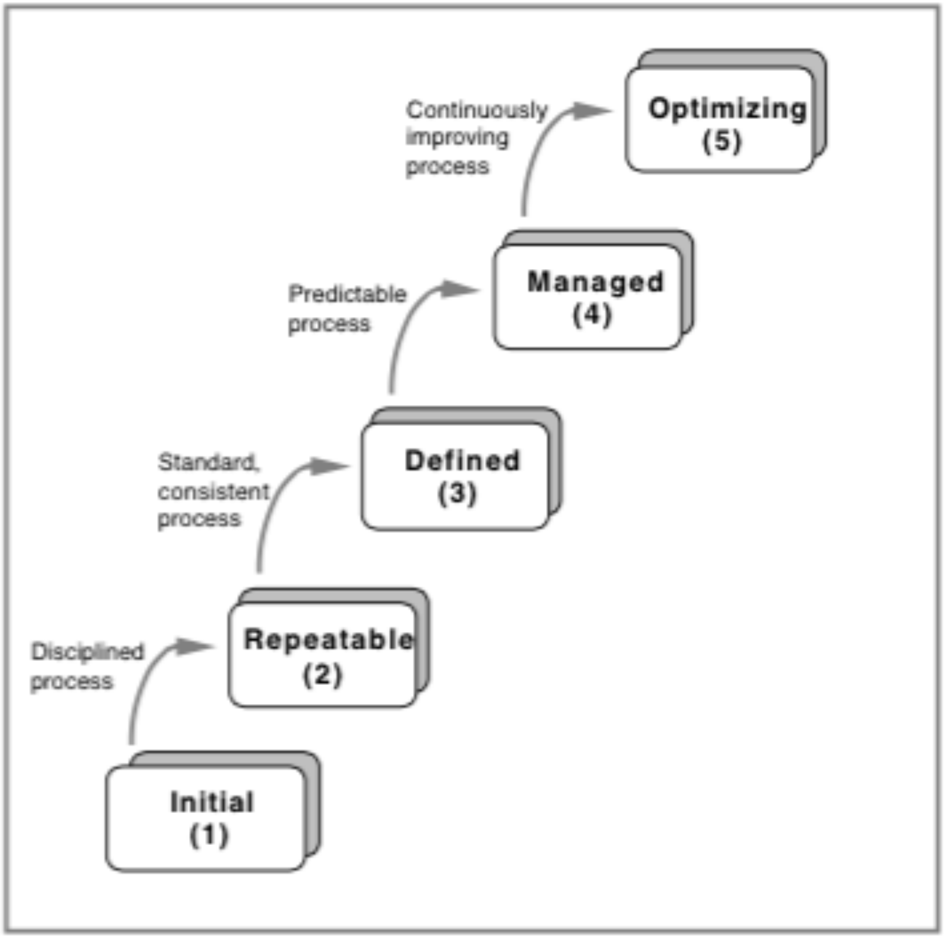


Figure 4 Capability Maturity Model [55]

The following data governance maturity models are attempts by key leaders in the field of data management to help identify the current state of organizations and provide them with a roadmap of how to achieve the next level.

Aitkens, et al.

In perhaps one of the more intensive studies in this area, from the period of 2000 to 2006 a survey of 175 organizations was conducted by Aiken, et al. that investigated the state of the practice of data management [16]. They analyzed the data management processes as derived by Burt Parker [56] in his work for the UD Department of Defense. This work effectively analyzed the individual disciplines within Data Management, delineating them by whether or not they were implementation focused, or strategic.

Table 1 Data Management Disciplines [56]

Process	Description	Focus	Data type
Data program coordination	Provide appropriate data management process and technological infrastructure	Direction	Program data: Descriptive propositions or observations needed to establish, document, sustain, control, and improve organizational data-oriented activities (such as vision, goals, policies, and metrics).
Organizational data integration	Achieve organizational sharing of appropriate data	Direction	Development data: Descriptive facts, propositions, or observations used to develop and document the structures and interrelationships of data (for example, data models, database designs, and specifications).
Data stewardship	Achieve business-entity subject area data integration	Direction and implementation	Stewardship data: Descriptive facts about data documenting semantics and syntax (such as name, definition, and format).
Data development	Achieve data sharing within a business area	Implementation	Business data: Facts and their constructs used to accomplish enterprise business activities (such as data elements, records, and files).
Data support operations	Provide reliable access to data	Implementation	
Data asset use	Leverage data in business activities	Implementation	

Subsequently, Aitken, et al. benchmarked the organizations against a maturity scale that was closely modeled after Paulk's work [55], with a focus on the use of data as an organizational asset.

Table 2 Data Management Maturity Scale, Aitken, et al. [16]

Level	Name	Practice	Quality and Results Predictability
1	**Initial**	The organization lacks the necessary processes for sustaining data management practices. Data management is characterized as ad hoc or chaotic.	The organization depends on entirely on individuals, with little or no corporate visibility into cost or performance, or even awareness of data management practices. There is variable quality, low results predictability, and little to no repeatability.
2	**Repeatable**	The organization might know where data management expertise exists internally and has some abili- ty to duplicate good practices and successes.	The organization exhibits variable quality with some predictability. The best individuals are assigned to critical projects to reduce risk and improve results.
3	**Defined**	The organization uses a set of defined processes, which are published for recommended use.	Good quality results within expected tolerances most of the time. The poorest individual performers improve toward the best performers, and the best performers achieve more leverage.
4	**Managed**	The organization statistically forecasts and directs data manage- ment, based on defined processes, selected cost, schedule, and customer satisfaction levels. The use of defined data management processes within the organization is required and monitored.	Reliability and predictability of results, such as the ability to deter- mine progress or six sigma versus three sigma measurability, is significantly improved.
5	**Optimizing**	The organization analyzes existing data management processes to determine whether they can be improved, makes changes in a controlled fashion, and reduces operating costs by improving current process performance or by introducing innovative services to maintain their competitive edge.	The organization achieves high levels of results certainty.

The results of the study yielded that only the area of Data Program Coordination was in a defined state. The other 4 areas of data management were either initial or repeatable. In this case, the results reflect Aitken's statement that organizations tend to regard data management activities as maintenance, rather than managing an asset.

Dataflux Data Governance Maturity Model [57]

As with most the models, the Dataflux Data Governance Data Maturity Model provides a means for organizations to identify where they reside on the continuum and an where the next phase is for them. The model is divided into 4 distinct stages with inverse dimensions depicting risk and reward. What is notable about this maturity model is that in addition to the four phases, parallels are drawn which relate the act of managing and governing data to the organization as a whole with terms such as "Think local, act global". This puts into perspective the tight coupling that managing data assets has with the organization and its culture.

Also interesting is that it identifies the types of technologies where data consolidation and integration often occur. Organizations will generally try to "piggy back" data clean up, integration, or reporting projects onto other more manageable initiatives. The bar between the second and third stages highlights the gap that is required to be crossed when moving to the "proactive" stage. This is generally where a change in how executive management views data is required. A commitment is required from management in terms of staffing, policies, tools, and the way projects are managed and governed. Some of the key aspects of the organization that are needed to be in place when moving to the higher stages of the model are:

- Data governance has executive-level sponsorship with direct CEO support.
- Business users take an active role in data strategy and delivery.
- A data quality or data governance group works directly with data stewards,

application developers and database administrators.

- Management understands and appreciates the role of data governance – and commits personnel and resources.
- Executive-level decision-makers begin to view data as a strategic asset
- Data stewards emerge as the primary implementers of data management strategy and work directly with cross-functional teams to enact data quality standards.
- A data stewardship group maintains corporate data definitions and business rules.
- Service-oriented architecture becomes the enterprise standard.

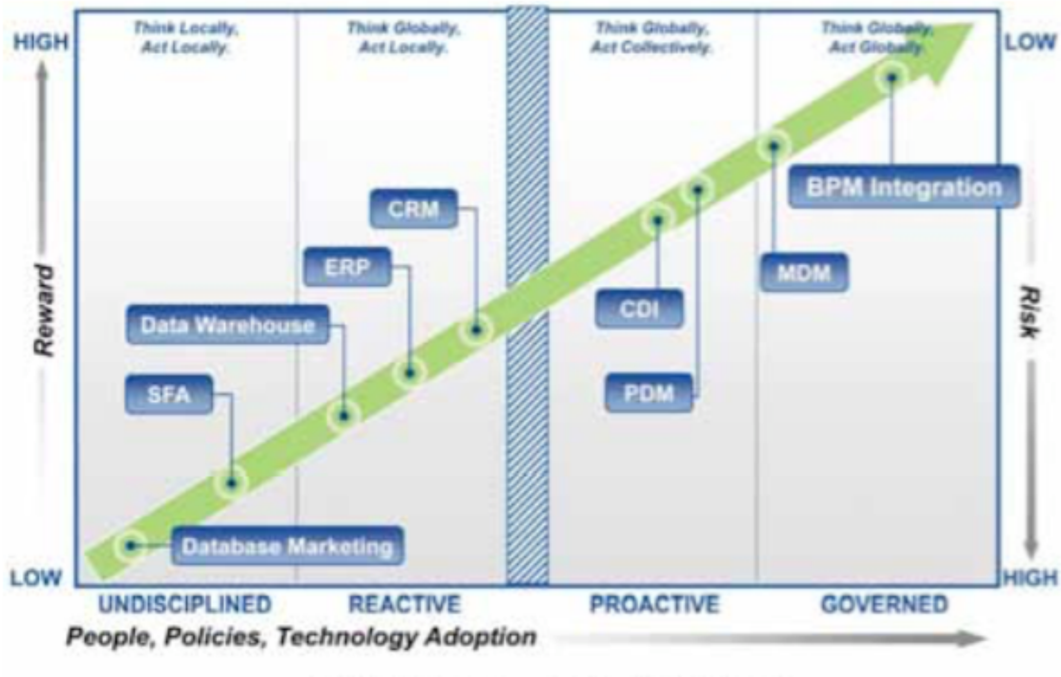

Think Locally, Act Locally.
Think Globally, Act Locally.
Think Globally, Act Collectively.
Think Globally, Act Globally.
HIGH
LOW
Reward
Risk
LOW
HIGH
CRM
ERP
Data Warehouse
SFA
Database Marketing
CDI
PDM
MDM
BPM Integration
UNDISCIPLINED
REACTIVE
PROACTIVE
GOVERNED
People, Policies, Technology Adoption

Figure 5 Dataflux Data Governance Model [58]

Gartner EIM Data Governance Maturity Model

In December 2008, Gartner introduced their enterprise information management maturity model [58]. As with other maturity models, it is to be thought as a continuum and evolving set of processes as opposed to a single effort. It is interesting to note that at the time the research report was written, Newman had predicted that over the next five years that organizations would start to manage information as a strategic asset within the organization. This can be seen in the next section where we look at the proliferation of data management tools that are currently on the market. Also of note is the term Enterprise Information Management. Here, Gartner's concept encompasses an integrated enterprise-wide approach to managing information where governance is a means to an end. The five primary goals of EIM are:

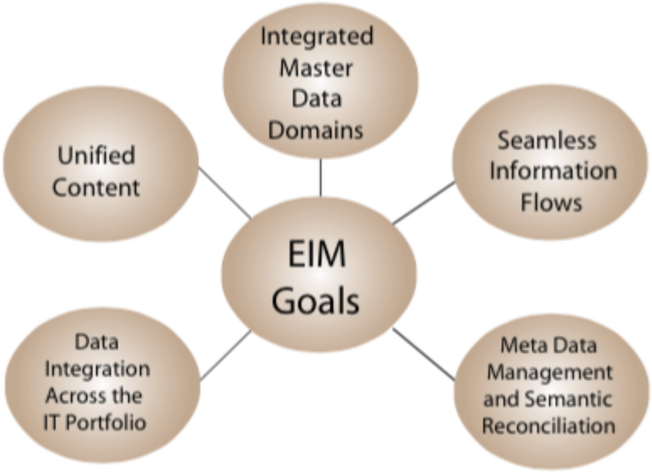

Integrated Master Data Domains
Unified Content
Seamless Information Flows
EIM Goals
Data Integration Across the IT Portfolio
Meta Data Management and Semantic Reconciliation

Figure 6 Gartner EIM Data Governance Goals [58]

In order to achieve the above goals, the following maturity levels have been established:

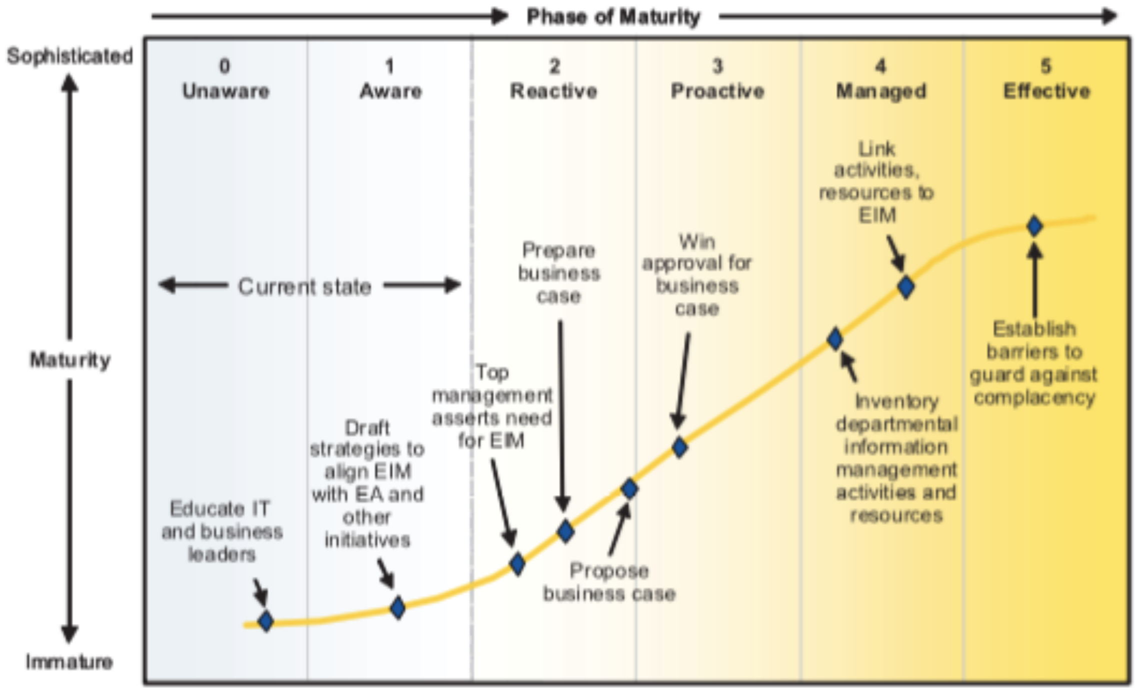

Phase of Maturity
Sophisticated
Maturity
Immature
0 Unaware
1 Aware
2 Reactive
3 Proactive
4 Managed
5 Effective
Current state
Educate IT and business leaders
Draft strategies to align EIM with EA and other initiatives
Top management asserts need for EIM
Prepare business case
Propose business case
Win approval for business case
Inventory departmental information management activities and resources
Link activities, resources to EIM
Establish barriers to guard against complacency

Figure 7 Gartner EIM Data Governance Maturity Model [58]

Table 3 Gartner Data Governance Maturity Levels [54]

0 Unaware	Strategic decision made without adequate information Lack of formal information architecture, principles, or process for sharing information Lack of information governance, security and accountability Lack of understanding of meta data, common taxonomies, vocabularies and data models
1 Aware	Understanding of the value of information Issues of data ownership Recognized need for common standards, methods and procedures Initial attempts at understanding risks associated with not properly managing information
2 Reactive	Business understands the value of information Information is shared on cross-functional projects Early steps toward cross-departmental data sharing Information quality addressed in reactive mode Many point to point interfaces Beginning to collect metrics that describe current state
3 Proactive	Information is viewed as necessary for improving performance Information sharing viewed as necessary for enabling enterprise wide initiatives. Enterprise information architecture provides guidance to EIM program Governance roles and structure becomes formalized Data governance integrated with systems development methodology
4 Managed	The enterprise understands information is critical Policies and standards are developed for achieving consistency. These policies and standards are understood throughout the enterprise Governance organization is in place to resolve issues related to cross-functional information management Valuation of information assets and productivity metrics are developed
5 Effective	Information value is harvested throughout the information supply chain Service level agreements are established Top management sees competitive advantage to be gained by properly exploiting information assets EIM strategies link to risk management, productivity targets EIM organization is formalized using one of several approaches similar to project management. The EIM organization coordinates activities across the enterprise

As with other maturity models, Gartner takes the organization from a state of chaos to control. It is interesting to note however that according to the figure above the action item of "Inventory departmental information management activities and resources" is placed at level 4 of maturity. Additionally, level 5 of maturity provides the benefit of "Information value is being harvested throughout the information supply chain". Major milestones such educating IT and business leaders, drafting strategies, getting top management on board all precede these critical steps. Although it presents as a "chicken and egg" problem, it is purpose of this research to show how a clearer view of the information supply chain, however dysfunctional it may be can be achieved with a lower barrier to entry and much earlier in the maturation process such that subsequent information management steps are easier to complete and higher levels of maturity are more attainable.

IBM Data Governance Council Data Governance Model

The IBM Data Governance Council, Data Governance Maturity Model [53] expands on the concepts of maturity phases and also includes 11 domains and the manner in which they interrelate. Established in 2007, this is one of the earlier models in which the interaction between the organization and the data lifecycle were specifically codified as elements that contribute to effective data governance. The 11 disciplines are segregated into 4 major groups: Outcomes, Enablers, Core Disciplines, and Supporting Disciplines.

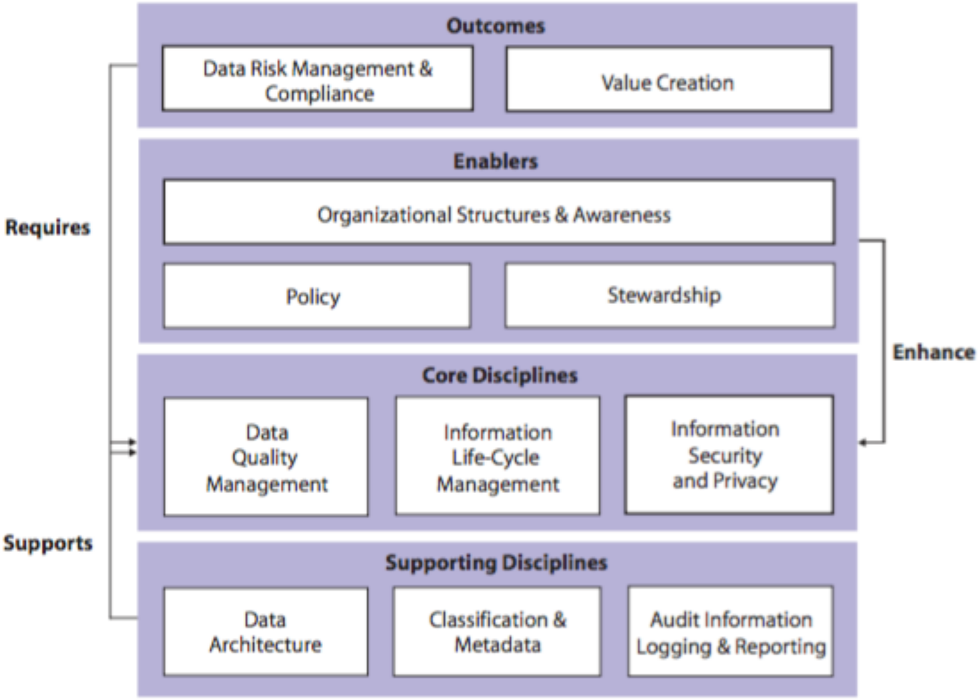

Outcomes
Data Risk Management & Compliance
Value Creation
Enablers
Organizational Structures & Awareness
Policy
Stewardship
Core Disciplines
Data Quality Management
Information Life-Cycle Management
Information Security and Privacy
Supporting Disciplines
Data Architecture
Classification & Metadata
Audit Information Logging & Reporting
Requires
Supports
Enhance

Figure 8 Elements of Effective Data Governance [53]

The above figure depicts the interdependency of the practices, organizations, elements, and policies required to effectively manage enterprise data. The following table depicts the 11 domain areas:

Table 4 IBM Data Governance Model - Data Governance Domains [53]

	Domain	Description
1	Data Risk Management & Compliance	The methodology by which risks are identified, qualified, and quantified, avoided, accepted, mitigated or transferred out.
2	Value Creation	The process by which data assets are qualified and quantified to enable the business to maximize the value created by data assets.
3	Organizational Structures & Awareness	Description of the level of mutual responsibility between the business and IT, and the recognition of the fiduciary responsibility to govern data at different levels of management.
4	Policy	A description of the desired organizational behavior(s).
5	Stewardship	A quality control discipline designed to ensure custodial care of data for asset enhancement, risk management, and organizational control.
6	Data Quality Management	Methods to measure, improve and certify the quality and integrity of production, test and archival data.

7	Information Lifecycle Management	A systematic policy-based approach to information collection, use, retention, and deletion.
8	Information Security & Privacy	The policies, practices and controls used by the organization to mitigate risk and protect data assets.
9	Data Architecture	The architectural design of structured and unstructured data systems and applications that enable data availability and distribution to appropriate users.
10	Classification & Metadata	The methods and tools used to create common semantic definitions for business and IT terms, data models, data types, and repositories. Metadata that bridge human and computer understanding.
11	Audit Information, Logging & Reporting	The organizational processes for monitoring and measuring the data value, risks, and efficacy of governance.

As the organization progresses through the stages of the IBM maturity model, different domains will become more relevant. This aspect of IBM's approach allows organizations to focus on particular areas of concern, providing next steps.

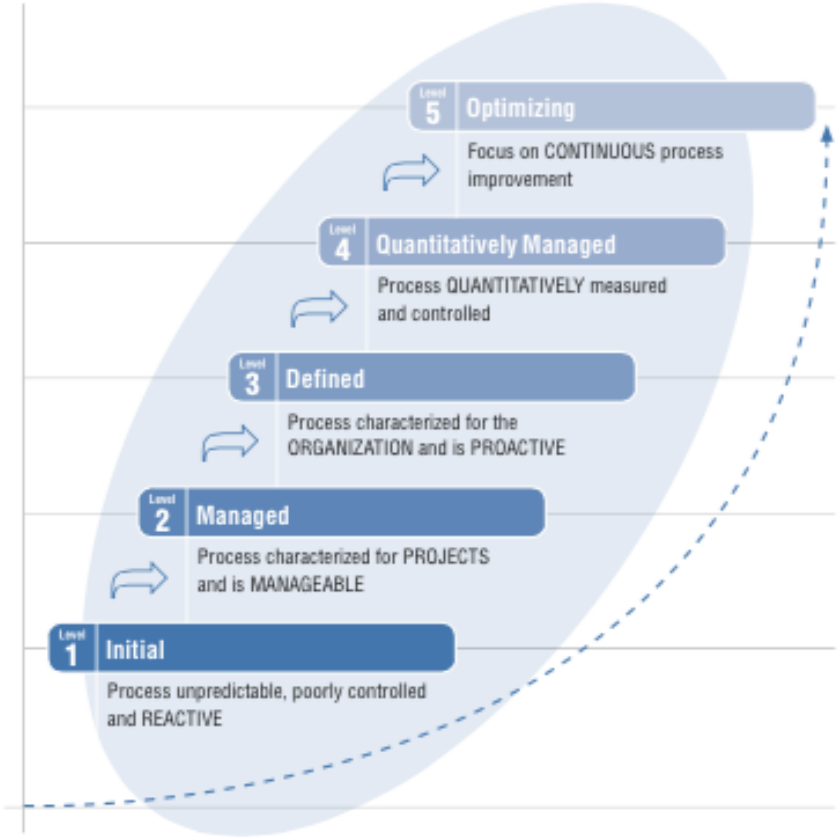

Level 5 Optimizing
Focus on CONTINUOUS process improvement
Level 4 Quantitatively Managed
Process QUANTITATIVELY measured and controlled
Level 3 Defined
Process characterized for the ORGANIZATION and is PROACTIVE
Level 2 Managed
Process characterized for PROJECTS and is MANAGEABLE
Level 1 Initial
Process unpredictable, poorly controlled and REACTIVE

Figure 9 IBM Data governance Maturity Model [53]

As with the Gartner maturity model however, the increased data awareness techniques in this research can assist in several of the domains specified above, such as: Organizational Structures Awareness, Stewardship, Information Lifecycle Management, Classification and Metadata, as well as Audit Information and Logging.

Oracle Data Governance Maturity Levels

As with the prior maturity models, The Oracle Data Governance Maturity Model takes a targeted approach at tacking enterprise data issues in phases. The key goals of data governance are [59]:

- To define, approve, and communicate data strategies, policies, standards, architecture procedures, and metrics.
- To track and enforce conformance to data policies, standards, architecture, and procedures.
- To sponsor, track, and oversee the delivery of data management projects and services.
- To manage and resolve data related issues.
- To understand and promote the value of data assets.

Achieving these goals is to be an incremental endeavor that increases scope across the organization, starting with project centric efforts, leading to cross-divisional initiatives.

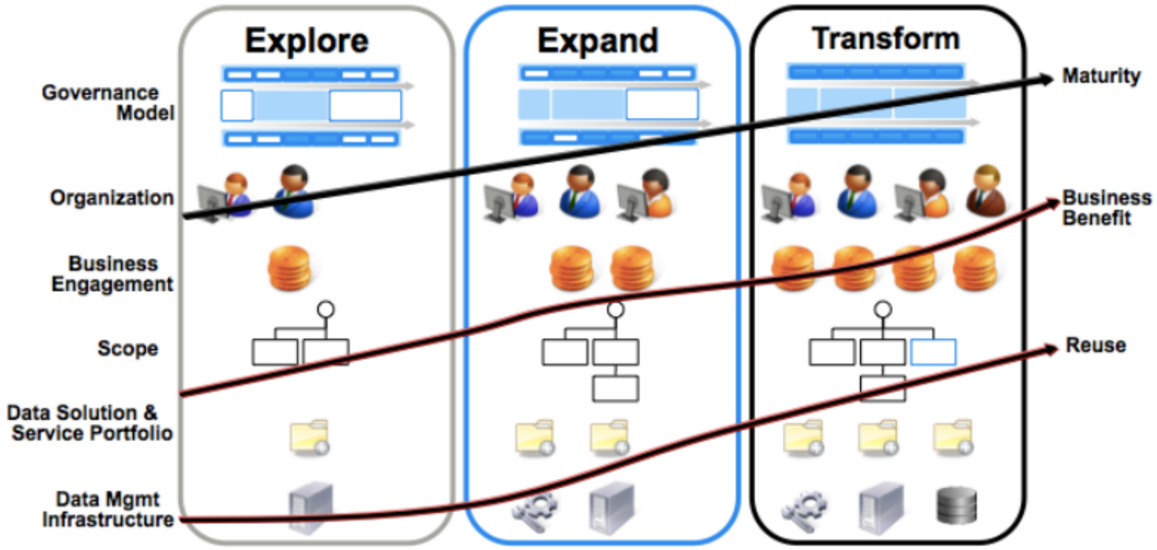


Figure 10 Oracle Data Governance Maturity Model [59]

In the above figure, we see once again, that maximum benefit to the organization is at the latest stages of maturity as governance initiatives spread from very localized approaches to more enterprise wide initiatives. This research however, will show that by starting with a low barrier to entry approach, such as a graph approach to increasing the awareness of data use throughout the organization, the entire enterprise is able to participate in low impact ways to contribute to overall enterprise data awareness.

2.2.3 Challenges to Effective Enterprise Information Management

Many of the issues with information management projects are that it's benefits tend to be difficult to quantify, making a business case for further investment in it hard to justify [16]. Some of this is due to the fact that management within organizations tends to make decisions on funding projects based on short-term results and trends. Enterprise wide data management initiatives, such as data warehouses however yield value when diverse cross functional information is brought within it [60]. As we have seen in the maturity models in the prior section, and also stated by Aiken [16], data management has had primarily a business-area focus rather than an enterprise one. The cross- functional nature of many of these projects adds an additional layer of complexity to manage towards success. Gil-Garcia et al [61] make the point that complex and sophisticated systems don't necessarily provide for efficient information sharing. Additionally, we see that knowledge workers in most organizations develop process at a local level, or "silos" as many organizational entities or agencies can at times take a defensive posture to protect their "turf" or in this case, their data [62]. In addition to creating additional integration issues at an implementation level, this local approach leads to variances in business definitions and concepts [63]. As data is maintained in different systems across the organization, we see that the efforts to integrate across multiple "gold copies" or definitive version of data adds to the confusion and degrades the accuracy of such data due to its different manifestations, while reducing the understandability of the data itself [7]. This rationalization effort around semantics is an area that has historically not been a significant part of the tenants of systems and process development, as they have focused more on localized transaction processing. It is the sharing and exchange of knowledge are what underpins much of the success of data management initiatives [64]. According to Das and Mishra regarding the practice of Master Data Management (MDM), "MDM is cross functional, it benefits from an organization that fosters collaboration between business and IT. Rapidly changing technology drive periodic application reengineering but the business customer remains with the organization. Clean consolidated and accurate master data seamlessly propagated throughout the enterprise can save millions of dollars, increase market base, improve customer loyalty and support sound corporate governance" [21].

Furthermore, data warehouse projects have a reputation of being costly endeavors, often missing deadlines, having unacceptable performance, and inevitably failing [65]. While a significant contributor to impediments to success is the fact that many organizations "have no idea of the enormity of the data they generate and are ignorant of data management methods being used" [66], we see from Elliot, et al. [26] that several studies point to weak executive sponsorship, organizational politics, lack of funding, and poor coordination with information technology staff and the user community as major obstacles to effective enterprise data management [25, 64, 67, 68].

Adding to some of the challenges faced with enterprise information management are that the commercial tools themselves are a relatively new, but albeit growing market segment. According to Gartner studies in both Data Quality and Data Integration tools [8, 69], we see that the markets for products in this space are showing substantial growth, reflecting the growing importance placed on these disciplines. While the major vendors in that are in the Magic Quadrant are well-established players in the information management space, such as IBM, Informatica, SAP, and SAS, it is also noted that common caution around their products is their cost and pricing models. High total cost of ownership and complex implementations are additional barriers to management embracing such tools. As Friedman states regarding data quality and data management, " As a discipline, it comprises much more than technology – it also includes roles and organizational structures, processes for monitoring, measuring, reporting, and remediating data quality issues, and links to broader information governance activities via data-quality-specific policies." [69], we see that technology's high cost to employ may not be commensurate with the part it plays in the greater discipline of enterprise information management. We therefore need to engage the enterprise to partake in the activities of data governance and stewardship, while removing cost and technology barriers to the process.

2.3 Linked Data and Ontology

2.3.1 Linked Data Background & RDF

The phrase *Linked Data* refers to a set of best practices and a collection of semantic web technologies such as RDF, OWL, SPARQL, etc… to allow users to publish, share, and link data on the world wide web or in the case of this research across corporate intranets. One of the primary aspects of *Linked Data* as opposed other means of storing data is that the relationships across the data are just as important as the data itself and are the very glue that allows for disparate data sets to be linked [70]. The main tenants of Linked Data, as defined by Tim Berners-Lee are [71, 72]:

- Use URIs as names for things
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information using standards.
- Include links to other URI's so that they can discover more things.

Here we see that as long as data is presented in this way and discoverable, it can be published and linked to any other data source on the web, or as mentioned within an enterprise intranet, so information can be shared across the organization. Another benefit about Linked Data is that when standards such as RDF are employed, the data becomes machine readable and explicitly defined [29]. This self-describing data, where it is completely separated from presentation and formatting allows publishers to not be constrained by a single vocabulary as relationships and links will allow the application to find a suitable definition. (While Lee and describes this as a benefit to Linked Data on the web, there are times, particularly within a “closed” or enterprise setting, where having a common vocabulary is beneficial.)

The machine readability, the use of HTTP protocol as a retrieval mechanism, combined with the flexible, link friendly graph structure of RDF allow for automated crawlers to comb through large sets of interlinked data and discover new data sources and their subsequent links [29]. It is this flexibility and ability to publish data that allows linked data to be a good fit for modeling enterprise data movement and metadata [73, 74].

The Resource Description Framework, (RDF) was developed as a means for describing metadata about resources by representing properties and the relationships that link them [70]. It is essentially the data model for linked data and was originally serialized based on XML. The relationships it models can be viewed or thought of in terms of graphs, where nodes pertain to resources on the web and properties and relationships are the edges. Essentially, an RDF statement describes two things and the relationship between them [75]. This relationship is also referred to as a *triple* as the basic structure is comprised of three segments: a subject, a predicate, and an object. The subject and object can are both Uniform Resource Identifier’s (URI) which represent a resource on the web (or intranet). A URI is “a compact sequence of characters that identifies an abstract or physical resource” [76]. The RDF definition also defines what data types pertain to what literals. It is important to note that RDF doesn’t require you to link to other data. It can very well describe entities and their relationships, however Linked Data, as defined by Tim Berns Lee requires the linking to other data sets that are described with RDF.

Resource Description Framework Schema (RDFS) is a means to define a vocabulary within RDF through the use of classes and the property *rdfs:type.* An RDF class is a grouping of RDF resources that are of a particular type of *rdfs:class*. Classes or resources can further be grouped through the *rdfs:subClassOf* property. The Web Ontology Language [77] which is the basis of the semantic web, is a vocabulary, in essence, and extension of RDF.

2.3.2 Ontology and OWL Background

As we have seen in the prior section, RDFS defines the data model for which we can create RDF statements. Much like a schema defines what rows can be entered in a relational database; the RDFS is the boundaries for what data can be put into the graph structure. To extend that further, OWL, is en extension of the RDF specification to increase the expressivity and inference that can be applied to various statements. The term *ontology* dates back to the 1700's and according to Webster's Dictionary, a branch of metaphysics concerned with the nature and relations of being [78]. In the Artificial Intelligence discipline, an ontology *is a formal specification of a conceptualization* [79]. Ontologies are based on classes and properties. Classes are definitions for the individual things that reside in them. Similar to object oriented programming, with ontologies, classes can also be thought of as ways to group commonalities and variations. For Example: A Ford is a type of automobile. In this case an object of "F-150", would be an *individual of* the class "Ford", which is a restriction of the class automobile. Properties can be thought of as the relationships between two objects. The relationships in turn can also be classes and extend from a superclass. For example, if we look at the following property: *is sister to*, we see that it relates two people in the following way Giavanna *is sister to* Gabriella. The property, *is sister to* can be though of as a restriction on a more general property such as is sibling to.

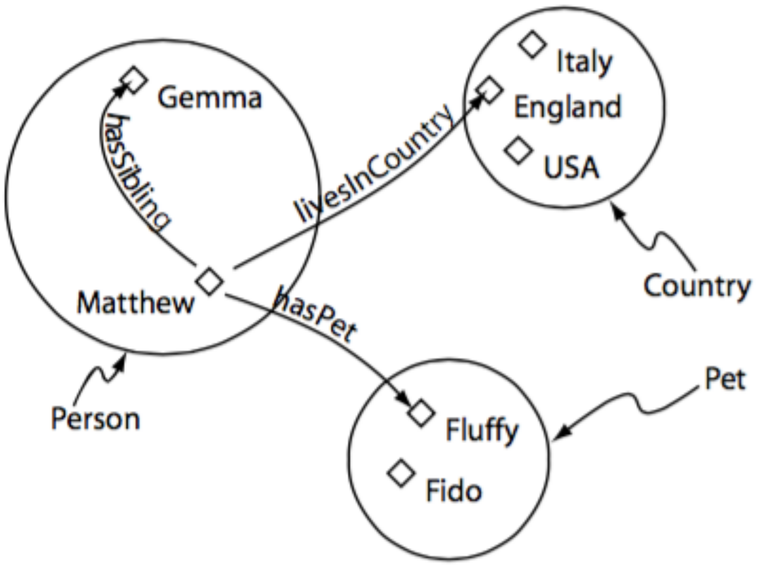


Figure 11 Ontology Example [80]

It is this richness and flexibility in specifying relationships that differentiate an ontology from a taxonomy, which is a hierarchical structure based on containment relationships. These relationships, called *object properties* are binary relationships that link 2 individuals together. Properties themselves can have characteristics, such as transitivity, which is seen in descendant style relationships, and symmetry, which implies a property, is the same in both directions. These relationships and their characteristics can be *asserted* at the time the ontology is constructed. The strength to OWL is when logic is applied to the ontology and additional relationships are *inferred*. Tools that scan the ontology and create these additional properties are called *inference engines* [81]. It is this process which creates new pathways that link individuals in ways that were unknown when the information was brought together separately. Along with object properties, which link two individuals, OWL also has data properties, which link individuals to data or literals. Dietz states that as enterprises have gotten more global, faster, and complex, there needs to be a way to describe them in a way that is concise, comprehensive, and essential. Thereby presenting the organization's essence, independent of implementation and realization. Ontology, with all it's expressivity is the manner in which to do this[82].

2.3.3 Linked Data and Ontology and it's Relation to Supply Chain Management and in Turn, Data Management

Just as we have seen enterprise technology expansion increase the need for heightened data awareness and governance, we are seeing an evolving distributed data supply chain that brings with it many of the fundamental concerns of Supply Chain Management (SCM). For instance, with manufactured goods, quality is judged by relying on knowing something's origin and how it was produced, it's *provenance* [83]. The term provenance, is used frequently in the antiquities and art community to describe an item's origin and travels through time to its current point. It is such a fundamental concept that a model for codifying it was established as The Open Provenance Model [10]. Data is being seen more and more in the same light of SCM. Supply chain management is a complex endeavor that involves a networked enterprise [14]. The network can expand across many different entities and systems. Much like in data governance situations, the hetrogentiy of supply chain management systems and the disunity of knowledge across the processes are major contributors to inefficiencies and inconsistencies [84]. Lu, et al. propose an ontology to help model and standardize concepts of supply chain management while giving insight into the interoperability problems of a networked enterprise. In this ontology, a common share model is being utilized to align the product with all the processes that have relationships with the supply chain. Lu states that "ontology is a promising means to unify the metadata model to express knowledge resources which are divers in types and disunited in forms"[84]. The manufacturing supply chain, much like data has many parties, processes, stages, and products that require the semantic richness of an ontology and the flexibility of linked data to model [85]. This thought is also reinforced by Levitin, Lee, Strong, and Wang as they propose that since the processes that create data have similarities to the processes that create physical products, data producing processes can be viewed as producing finished, or intermediate data products for consumers, which can be individuals, or other systems[86-89].

2.4 Conclusion

As we have seen in this chapter, there is a large body of work that has described the importance of data quality across the enterprise. The lack of quality data is directly attributable to the manner in which it is governed. We then looked at various governance models and explored the issues related to implementing proper data governance across the organizations and see that good awareness of data and how it relates to processes in an organization can lead to better data governance.

We then explored graph databases, resource definition framework, ontology, and the web ontology language. These flexible and expressive methods of modeling knowledge do well when modeling complex linked processes with many touch points. Two such instances are supply chain management, and by extension because of it's similar manner, corporate data flows. It is established that poor awareness of enterprise data decreases the ability to govern and manage enterprise data [5,12, 13]. "Without a good understanding of the relative size of the data challenge (degree/complexity for a specific scope) facing your organization, investments cannot be justified" [32]. This in turn increases data quality issues through latency, using the wrong definitions, increasing time to derive impact of issues, etc. in organizations can put it at risk and hinder efficient operations.

This conclusion shows that flexible and cost effective tools such as those provided by graph data structures and ontology can help support adoption and participation and in turn improve overall data awareness, thereby improving corporate data governance efforts, which will have a positive impact on corporate data quality.

Chapter 3

Solution Methodology

3.1 Introduction

As we have seen in chapter 1, poor data quality, resulting in a lack of stewardship, can put organizations at risk, create inefficiencies and waste, and cause opportunities to be missed. Increasing visibility to corporate data can mitigate the effects of some of those actions, by eliminating them altogether or providing the necessary knowledge to address their impact quickly. In this chapter, we will present a solution to which will help mitigate the difficulties in enterprise data governance by providing organizations with a low cost, low impact means to organize their information assets, processes, and initiatives. To illustrate our solution, we will be using a use case that centers on "Acme Corporation". Acme Corp, is based on corporate structures and enterprise data architectures that were researched for this dissertation. It represents a large organization with many disparate systems and processes across many functions and geographies. Many of the systems exchange data across the organization and are central to enterprise processes and workflows. Various individuals and departments play various roles in e ownership and management of that data, from owner, to consumer.

This solution is comprised of several components such as an ontology, graph browsing tool, a crawling mechanism, and query/reporting constructs which will allow users across an organization to collect, update, and report on information interactions across the enterprise. This information will allow for more effective data governance, as more people are able to take part in building greater data awareness.

3.2 Specific Use Case For Data Awareness

One of the most common instances where data awareness assist in better data governance is in the identification of an impact across the enterprise should there be a change to key data assets. For example, Acme Corp. had to change the field length of their employee ID in their core HR system of record, PeopleSoft. As that HRIS system had been in place for over ten years prior to this change, a great deal of other systems had come to rely on employee data coming from it. Those systems in turn sent data to others, or were a source for reporting tools. The effort involved in identifying all of the systems, their information flows, and the processes that would be impacted by the change was a major undertaking.

In order for data impact analysis to be viable in the enterprise, it needs to be adopted by as many people as possible. As we have seen from the research in chapter 2, highly centralized enterprise data initiatives tend to get derailed due to low adoption, high cost, and a long time before ROI is realized. Additionally, a solution must be able to rationalize data derivations and infer impacts. For example many organizations have a "mentorship" program in which an employee serves as a mentor to another one. In those cases the "mentor ID" field is the same as the employee ID field but shows the employee ID operating in another capacity. In another example, the Sales department can get a file of employees and load them into their systems and load the employee ID data in the Salesperson ID field and then send that data to other users as well. It is these "mutations" that must be captures to get the full picture of data throughout the organization.

In this use case, we will present the way in which our solution provides a way for key stakeholders to organize, store, and view this information in a scalable manner.

3.3 Solution Overview

The solution that is being presented is based on the reality that enterprise data centric projects, have a high failure rate and require the participation of many people within the organization [90-92]. Furthermore, reducing the costs associated with these projects and the proprietary systems eliminates barriers to entry and increases the adoption throughout the organization. To that end, the solution is to use a graph based data store in the form of RDF, or inked data, along with semantic web techniques, such as ontology, OWL, and SPARQL to inventory corporate data assets, and model how the data in an organization relates to people, processes, and concepts.

The core ontology for which this solution is based upon is built to capture key concepts in data governance, information management [4], as well as organizational structure and key processes. It forms the vocabulary and framework for an RDF data store to be based upon. Having a common vocabulary allows the organization to maintain consistency in how data is represented. Furthermore, the graph based nature of RDF drives two major aspects of the solution: flexibility in that participants can provide data at varying levels of detail, and transitivity, in that long chains of data movement and lineage can be described in a relatively simple manner.

In addition to the ontology and graph based data store, other components are proposed, such as an enterprise agent, with the goal of keeping the graph in sync with changing enterprise data assets. Also, various graph viewers and open source modeling tools are proposed as means to view and update aspects of the ontology and graph. Finally, sample SPARQL queries are provided in order to show the richness of the solution in terms of providing data across many relationships in a more streamlined manner than a relational data store.

3.4 Comprehensive Solution Design

3.4.1 Linked Data and Ontology is a Suitable Foundation to this Problem.

Linked data is well suited to the task of improving corporate data quality through increased meta-data visibility because of the flexibility and expressiveness of semantic web technologies and graph data structures. As we have seen, the meta-data visibility required to effectively improve data quality goes well beyond describing the actual structure of the data (databases, rows, columns, etc.). Comprehensive meta-data visibility also entails how corporate data is being used and distributed across the organization, by whom, and what business concepts or enterprise process does that data pertain to. The ability to model those complex relationships across the organization in a standardized manner that can be shared is necessary to describe how data interacts with the organization.

The key attributes of linked data and ontology that make is suitable for increasing meta-data visibility are:

- Expressiveness: Complex and rich relationships between people, process, and data can be modeled in the form of triples that in turn comprise a graph structure. These relationships can link data to other entities, both physical and conceptual throughout the organization thereby creating a graph structure that captures how data, systems, and people integrate. By extending RDF with OWL, Web Ontology Language, complex rules and relationships can be established in the form of ontologies that provide the basis for inference engines to traverse the graph and assist in drawing further conclusions from the data. The additional benefit is that the relationships, allow the data to be “self-describing” hence tools that process the graph information can be light weight and needn’t be concerned with maintaining relationship or rules.

- Flexibility: Another benefit that semantic web technologies provide is the flexibility to specify data at varying levels of detail [93]. One of the benefits of being a graph structure is that there is no root. Unlike XML, which is a tree-based structure, in RDF, no one resource is of any inherent significance to another. Therefore it is easier to augment a graph with new data, as it is as if it is being placed next to it. The semantic web is based on an open world assumption in where data can be specified about anything and it is assumed true until contradicted [30]. While this can pose an issue in the world wide web, when it is being used in a closed environment, such as an organization controls can be put in place in the forms of rule, assertions, and axioms that can keep standardization while allowing a high degree of flexibility. Flexibility is very important when dealing with real-world aspects such as adopting such a solution across the enterprise, as well as maintaining it. Allowing teams to enter information as they have it, at varying levels of detail, or just as much as they need to will foster easier adoption. Systems that are based on relational models can encounter issues when dealing with data at varying levels of detail, causing issues with parent-child relationships, and “orphaned” rows. Conversely, being able to maintain a standardized framework to house corporate meta-data provides a very powerful data source that can be used enterprise-wide for providing both high level, strategic views of meta data as well as the ability to provide tactical information to support operational decisions, that directly relates to the lowest level of detail (i.e. tables and columns)

- Ease of publishing**:** As stated above, regarding flexibility, the open world assumption allows people to publish anything about anything regarding in the semantic web very easily, by just exposing a URI. Semantic web browsers will, when seeing a reference will access that resource and integrate it as part of the current view. Reasoning can also be applied to those structures as well thereby allowing full integration across different semantic data providers. The fact that it is relatively easy to expose information to the semantic web (in this case a semantic web that is internal to an organization) allows many parties to easily provide data for sharing, thereby, providing a “crowd-sourced” approach to building a knowledge base. This is a critical element that makes it well suited for meta-data management. Large-scale meta-data management efforts require participation across the organization to be successful. In those cases, not only can the content of the knowledge base grow incrementally, in an agile form, but the data that is provided will be exposed to many more parties across the organization so information that may be missed if this were a traditional centralized approach, being headed up by one team, has a higher chance of being captured and corrected. For example, one very large team may think they know all of the recipients that they send their data to in the form of messages or feeds. Upon publishing this information, another team, which may be a data recipient, may see that they are not included as a dependency. They can therefore take corrective action. Had this been a traditional two party relationship, with a centralized team, they may not know of the additional data dependencies, or it may have taken longer for the dependent teams to be made aware of the omission.

- Inference**:** Inference is one of the main benefits of OWL [30, 81]. Here, a reasoning engine will infer new data based on the assertions and rules within the knowledge base. This ability to infer data is necessary for the enterprise data graph to maintain its flexibility. In the case of meta-data management, it is a necessary that teams be able to enter the data at a level, which is easiest for them to keep up to date. Inference is very important because it will be able to span "gaps" in the data. For example a team may only submit that they have a particular system that is being accessed by multiple individuals. These individuals can range from stakeholders to technical owners. In another case, they may only know that a given department is accessing their system. Inference will draw the appropriate conclusions as to what departments are stakeholders and technical owners based on where in the organization those individuals are. This can be extremely important to management, as they may need to only see information at more summary levels. Furthermore, there are several open source inference engines that can be used for this purpose [94-97]. Similarly, inference can negotiate varying levels of detail with systems information as well. Ideally, table and column level information would be present in the data graph so a true inventory of low-level meta-data exists. However, to get started, that may not be possible, or it may be too difficult for some groups to maintain. Reasoning engines therefore can create the appropriate entailments, or implied facts that will be able to link organizational information to system information, both at varying levels of detail. From a practicality standpoint, the amount of information that needs to be loaded into the ontology, along with the effort to do so is significantly reduced as much of it can be derived. This furthermore contributes to lowering the barrier to entry in fostering adoption across teams.

- Rich frameworks & tools: Semantic web technologies also offer a number of mature, open source frameworks, tools, triple stores, reasoning engines, visualizers, and wikis [98, 99]. In addition, similar to RDBMS' there is a robust query language called SPARQL, which allows for querying the knowledge base. Given the graph based structure of ontologies, they natively lend themselves to visualization and interactive querying [33]. In this case, the ability to navigate through a graph of an organization and readily see data movement and dependencies is not only much more intuitive than a spreadsheet, but allows many more people in the organization to have a greater awareness of corporate data [93]. Furthermore, from a system integration standpoint, frameworks from Java to C to Visual Basic exist that provide "endpoints", or services that provide storage and retrieval, inference, querying, and manipulations functions programmatically, for integration into custom applications. Finally, there exists several types of RDF stores, both open source and proprietary that are available for handling large numbers of triples efficiently.

The combination of rich open source tools, and the low cost of entry from an effort standpoint make semantic web technologies an attractive alternative to higher cost, proprietary, centralized approaches to metadata management. Additionally enterprise wide meta-data projects rely on a high rate of participation, therefore the flexibility and open access to such a solution fosters an agile approach to enterprise meta-data management as well as higher quality input as many more individuals across the firm can visualize the inter dependencies of systems, data, and people.

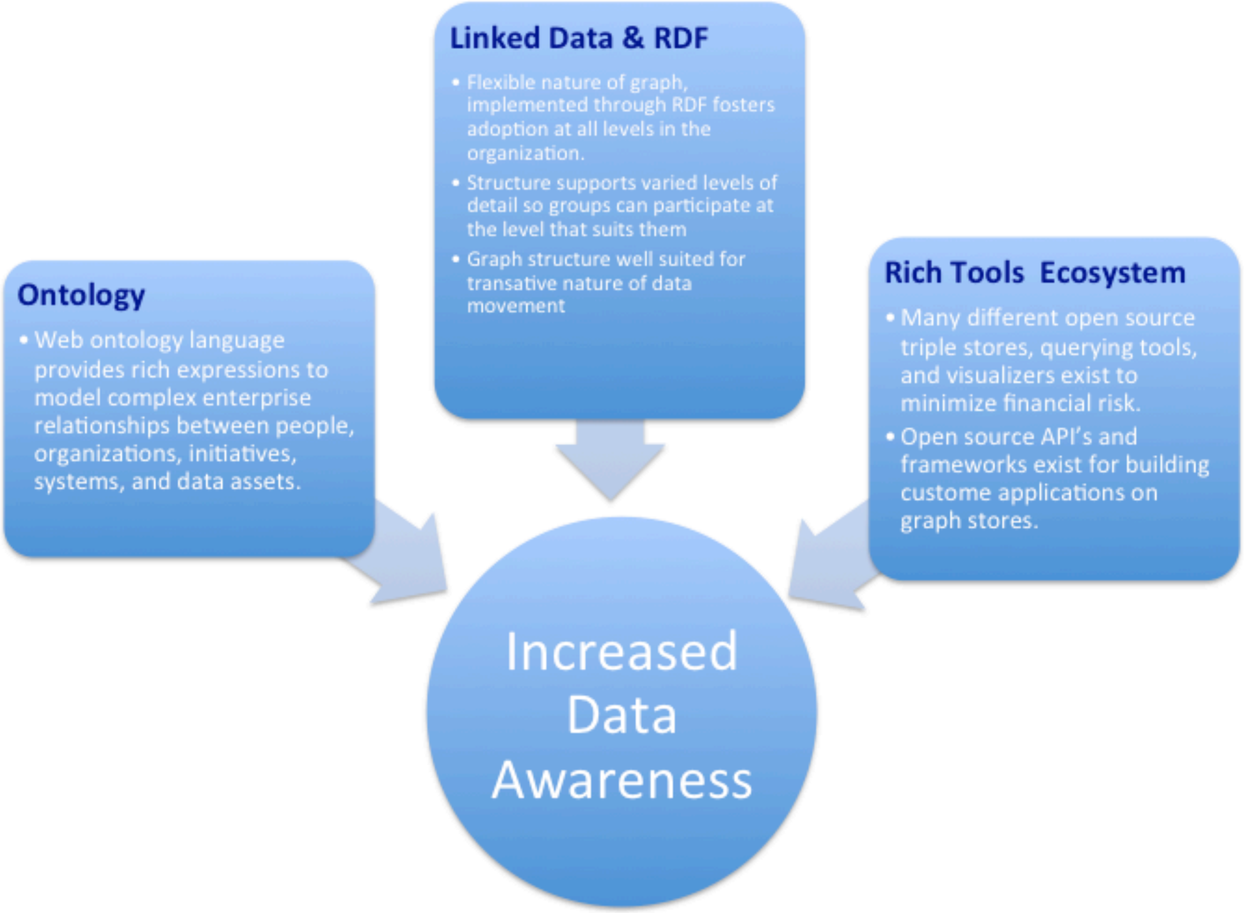

Linked Data & RDF
• Flexible nature of graph, implemented through RDF fosters adoption at all levels in the organization.
• Structure supports varied levels of detail so groups can participate at the level that suits them
• Graph structure well suited for transative nature of data movement
Ontology
• Web ontology language provides rich expressions to model complex enterprise relationships between people, organizations, initiatives, systems, and data assets.
Rich Tools Ecosystem
• Many different open source triple stores, querying tools, and visualizers exist to minimize financial risk.
• Open source API's and frameworks exist for building custome applications on graph stores.
Increased Data Awareness

Figure 12 Overall Solution Approach

3.4.2 Solution Requirements

In order to effectively address the obstacles that stand in the way of effective data governance, a solution that improve corporate data awareness, must include the following function points:

- A common vocabulary in which corporate data assets, business concepts, processes, data governance information can be described.

- A means to catalog and present system meta-data at varying levels of detail, so people throughout the organization can see what data is available and what exists in areas beyond their own.

- A means to associate system meta-data to all appropriate parties within the organization involved throughout its lifespan. Such visibility will assist in viewing and sharing ownership stewardship information, while also seeing the breadth of stakeholders of data as it moves from system to system.

- A means to catalog and present the movement of data from one party to another in order to see the reach of a piece of data throughout the organization. This knowledge is crucial in order to perform impact analysis, effective change management, or to monitor data consumption across the enterprise.

- A means to associate corporate business concepts to the actual data that comprises it. As organizations have many abstract concepts that are shared across it, it is crucial to identify the core data and systems that comprise those concepts.

- A means to query and report on the information contained in the framework.

In addition to the above items, such a solution must be presented to the organization in a manner where there are low barriers to entry; it's unobtrusive, flexible, economically viable, and maintainable. This is required because such a solution is essentially reliant on maximum participation throughout the organization in order to fully realize the relationships between people and the systems. Additionally, the value of any modeling solution is that it accurately reflects the real world objects it presents, hence maintenance and upkeep are a must. Active maintenance is done by empowering each group to publish their own information in an agile fashion at the level of detail that suits them as they all have “day jobs”. With time, and participation the organization will have a very rich and current meta-data store where, a clear view of corporate data and it’s movement and involved parties can be readily viewed.

3.4.3 Solution Architecture Overview

At the heart of the overall architecture is a core ontology that is the basis for a graph data store and holds information for core meta-data throughout the organization. That core graph (*Enterprise Data Graph*) and its ontology is then leveraged by other groups in their graphs (*Departmental Data Graphs*) throughout the firm and extended with additional information that pertains to their data structures and organizations. Much like any data modeling tool, keeping the model, in this case the graph structures current and consistent is crucial to having it accurately reflect the organization. The following illustration shows a high level view of the components:

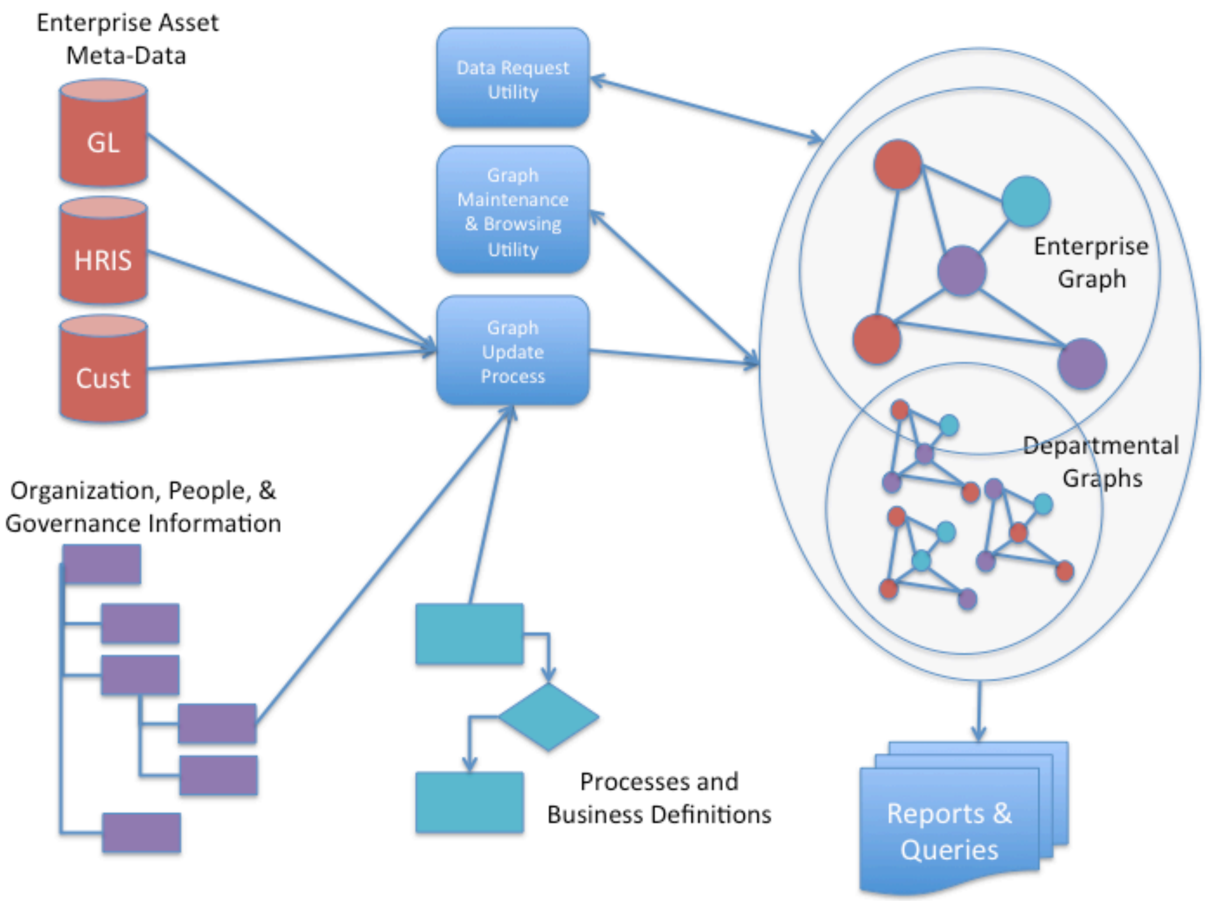

Enterprise Asset Meta-Data
GL
HRIS
Cust
Data Request Utility
Graph Maintenance & Browsing Utility
Graph Update Process
Enterprise Graph
Departmental Graphs
Organization, People, & Governance Information
Processes and Business Definitions
Reports & Queries

Figure 13 Solution Components

Key components to the solution are:

- Enterprise Graph: The Core data graph is the heart of the solution. It is an ontology that contains enterprise wide reference or master data, such as organizational information like departments, divisions, and people, regulatory entities. It will also contain information describing the relationships that depict data movements and corporate initiatives, such as provenance information, project related information, and business domain information. In addition to particular entities that comprise these classes, as with all ontologies, the relationships that link them are included as well.

This core data graph will then import the various departmental data graphs via the IMPORT directive in OWL:

`<Import>http://www.owl-ontologies.com//NEW_ONTOLOGY</Import>`, thereby bringing together information from across the enterprise.

- Departmental Graph: The Departmental Graph contains localized information for a particular department or other atomic working group. The purpose of such localized graphs is to encourage enterprise wide participation in publishing information regarding data use and stewardship by breaking up the task and putting the responsibility in the hands of people that work with their own data on a regular basis. The graph can contain as much or as little data as necessary to describe the local environment and contribute to the enterprise-wide goals. Particular information that a team would best "own" includes: physical and infrastructure related information about the data, stewardship information, provenance information regarding data movements, and abstraction information, such a business domain object mappings. The departmental graphs would in turn import from as well as contribute to the core data graph which will provide access to all the other departmental graphs throughout the firm, as well as providing a common "hub" for enterprise wide meta data.

- Enterprise Asset Meta-Data: These are the actual technology assets within the organization, such as systems and the databases they reference. These are the elements for which we want to capture and track their movement, usage, and governance throughout their lifespan.

- Organizational, People, & Governance Information: These are representations of the organizations structure, such as divisions, departments, legal entities, and people. This information forms the basis of the stewardship and governance relationships that bring people and data together.

- Process and Business Definitions: This information pertains to how the technology assets relate to the business. Most technology solutions are part of a business process that is in turn “owned” by one or several functional areas. An example of this is the Recruiting Process at large organizations in which the process is “owned” by the Recruiting department but has involvement from other areas within HR. Having the functional processes related to key data elements is imperative in identifying impacts to changing data structures such as in upgrades, or conversions. Additionally, business definitions are the terms used in organizations that can pertain to entities, metrics, attributes, etc.. Many times these are formulae that refer back to supporting data or are terms used by the business. Problems arise when there is confusion over the meaning and the supporting data is not clearly identified. As an example, in HR, the term Headcount can often be construed to mean all people currently working, or employees currently working, or all people that are currently on payroll (e.g. including people on maternity leave). This information will be essential to linking data to their appropriate definitions and their owners in one place.

- Graph Update Process: This process encompasses both manual and automated ways to continuously update the graph based on the state of the organization. In it’s most robust form, the update process can be fully dynamic, much like an agent, or a crawler and update the graphs based on the data base structures, organizational entities, and business processes. It can however be manual as well in which there is operator intervention to extract the information and load into the graphs. While this approach will be somewhat limited in large organizations with complex hierarchies and many systems, the flexibility of linked data can allow for several approaches. We will see several ways in which to update the graphs in chapter 4.

- Graph Maintenance & Browsing Utility: This process allows the user to view the data graphs to observe view the data landscape within the organization, but also make updates, as needed and are permitted to. Most of the manual updates related to processes and definitions would be carried out via this mechanism, however, departments may also be able to update their own graphs as per their organizational needs. We will look at several ways in which to update the graph in Chapter 4.

- Data Request Utility: One of the key aspects in this solution is being able to track the lineage, or provenance of data throughout the organization. In order to capture the data needs within the organization, a utility or process needs to be put in place where these requests can be captured and then if and when they are realized, the data transmission (data feed, service call, etc.) will be properly aligned to it's upstream sources, and downstream dependencies, as well as its owner. Here, all data requests would be entered through a tool, which would capture the information that is sought, and the requesting individuals, as well as the target for this data. The user would be able to brows the existing landscape of data and select the fields they would like. Not only does this approach log how is getting what data for visibility into stewardship, it also forms the basis of creating provenance information so true data lineage can be tracked. Such a tool can also be employed to compel data recipients to re-certify that they still require the data over time so as to minimize “dead” data feeds. Additional data capture tools can be used to map applications to various project initiatives, or to map data objects to business domains that are utilized by subscribing applications.

- Reports & Queries: These components allow the organization to extract information from the graphs in either a standardized or interactive way. There several ways in which this information can be viewed, the primary way being SPARQL, which is a query language for graph databases. Additionally, graph visualization tools, such as Protégé (see below) can be used to extract and or view information. These will be reviewed in the next chapter.

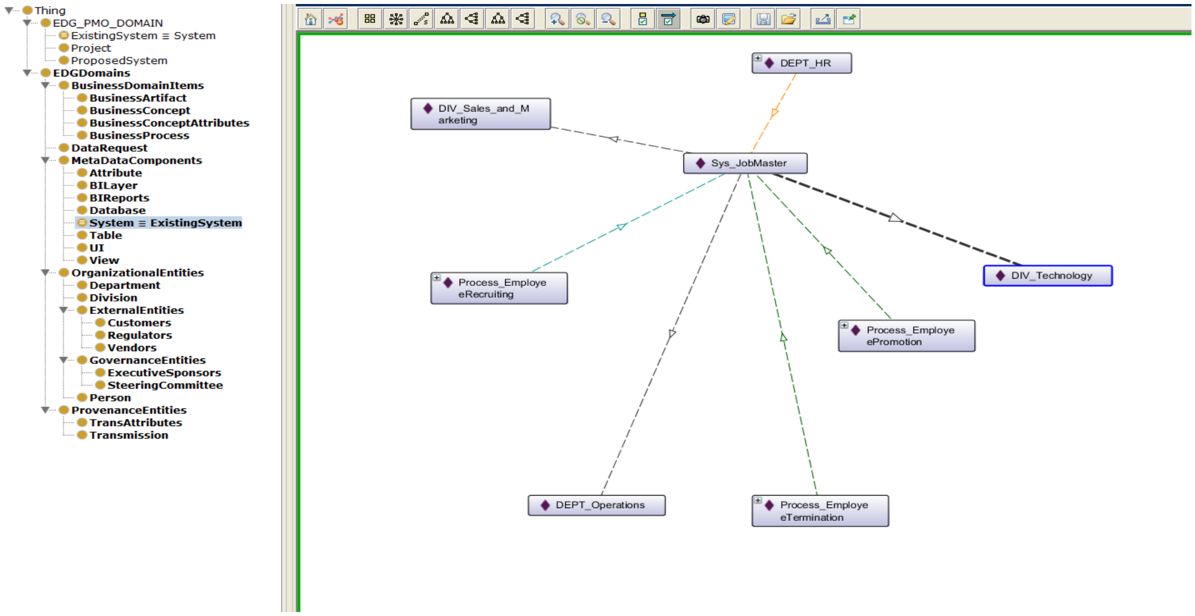


Figure 14 Using Protégé to Browse an Ontology

The actors involved in the process will be both technology and functional personal that have an understanding regarding their data movement. As the system matures, graphical interfaces will allow business people to view what data they have available to them throughout the firm and how it is used and propagated. Most importantly however, this increased exposure to corporate meta-data will aid in data governance as people and departments are clearly identified as creators, consumers, and stewards of enterprise data. Once this landscape is in place governing policies can designed and implemented.

3.4.4 Ontology Classes

At the center of the enterprise data graph is an ontology, which is the core model for which the enterprise data graph will be built upon. The ontology itself will model well-known concepts within the organization, such as department hierarchies, system components, people, committees, etc. This will allow for greater standardization across the organization. Business concepts can also be modeled as well and linked appropriately to the underlying data and teams that support them. Business concepts will contain the most semantic diversity across the firm as many groups can claim ownership of the same concept. Ideally, the act of aligning multiple business concepts can in itself be a valuable exercise and resolve to a deterministic conclusion. These classes, in conjunction with assertions and rules, comprise a vocabulary for which data, and its interaction with the corporate entity can be reflected. One thing to note is that much of the ontology is based upon containment as opposed to inheritance. To overcome this, appropriate naming conventions and object properties were established to model those relationships.

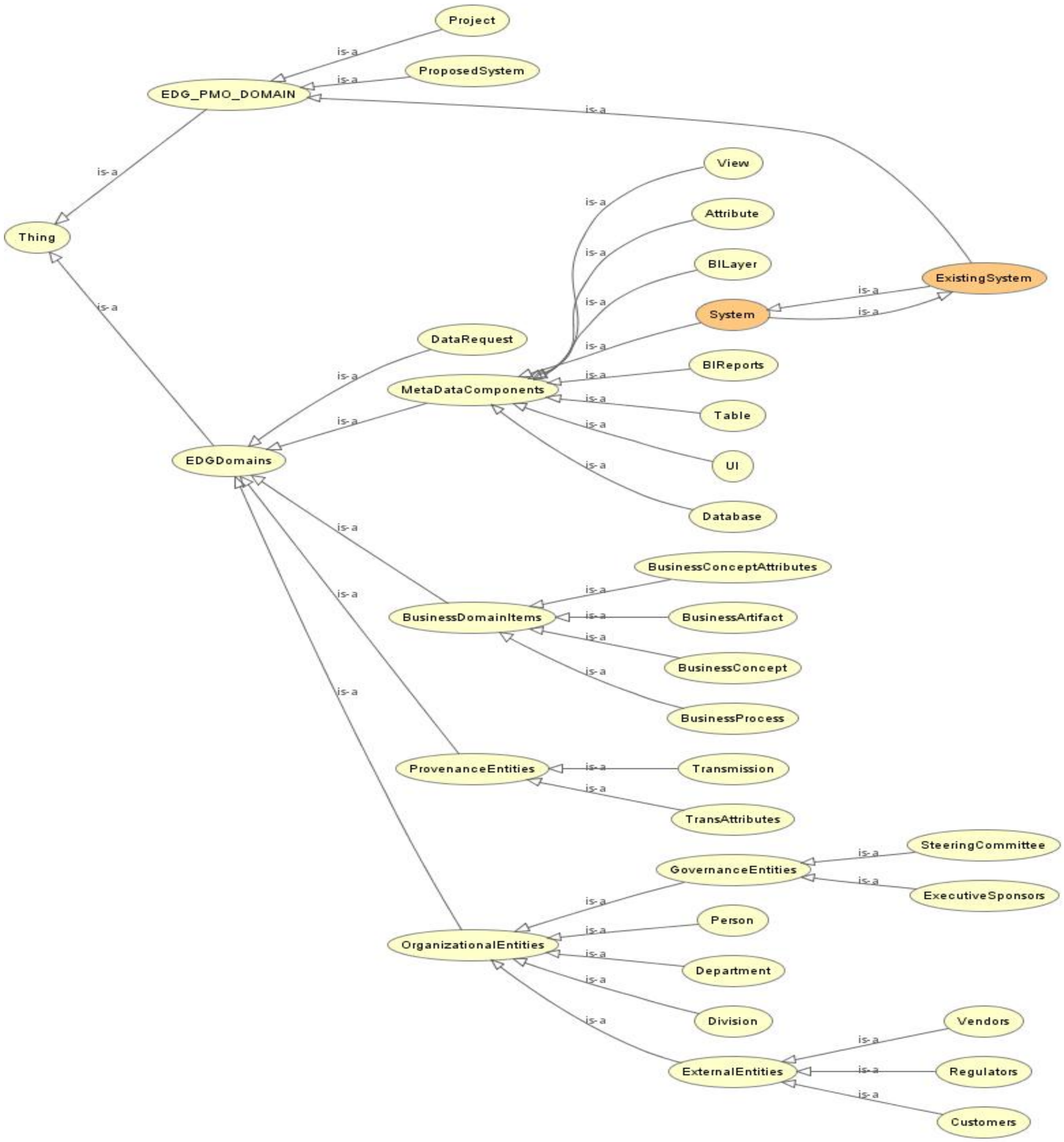


Figure 15 Ontology Classes

We will now proceed to define the key classes within the ontology:

Organizational Entities

The intent of this class is to model the organizational structure within a corporate entity. In many organizations, there are multiple ways to organize those entities, such as functionally, with divisions, such as Finance, Marketing, and Information Technology, or geographically, with regions, cities, offices, etc., or product centric, based on what a collection of entities produces. Regardless, semantic web allows for any number of different organizational structures. It is important to note that as many of these structures are based on composition, rather than inheritance, appropriate properties will need to establish to show the appropriate ordering. In most cases however, despite the various different ways to establish an organizational "tree", the end product, or the "leaf" is usually people. People are an important part of the meta-data graph, as assigning personal responsibility is critical to ensuring accountability and active data stewardship. In many cases, only key individuals need be known, such as "technical owner", or "business sponsor" and groups, such as stakeholders can resolve to organizational entities, such as departments, regions, or divisions. The ability for inference is very important for browsers of the graph to navigate the corporate hierarchy, as various levels of detail can be present.

Meta-Data Entities

Meta-data components are the items that reflect the actual technical artifacts within the organization. As stated earlier, the data persistence format that this research is focused on is structured data in a tabular format, which constitutes primarily RDBMS's and files in a tabular format, such as .CSV or fixed width. At the highest level there is a system, which is a collection of processes and technologies to support a business process. The system is comprised of various components, such as user interfaces and databases. Databases are subsequently comprised of entities, which in turn contain attributes. Through the use of object properties, meta-data information is linked to both organizational and data provenance classes in order to model how data is being used by whom and throughout its lifespan. Additionally, the business concept class is also linked to these classes in order to identify where the persistence resides for abstract business concepts. For example, there may be a business concept called “worker” which subsumes three entities: an employee table, a consultant table, and a temporary help table. In some cases knowing the tables alone can be sufficient to provide appropriate visibility on meta-data, hence columnar information, while helpful, may not be necessary, easing the effort on the team to keep the model current.

The meta-data components class will require the attention from various teams that are populating the ontology with their particular information. In order for it to accurately reflect the real world, it needs to be kept up to date, which can be done both in an automated fashion with agents that comb through database catalogs or control files, or as part of the software development process within the organization. This will be discussed further in Chapter 4.

Provenance Entities

Data provenance refers to the lineage of data throughout its lifespan. This includes not only how and where it was created, but its movement and transformation as it gets disseminated to other systems for their use to its eventual end state. This class is centered on capturing concepts of data “transmissions” from one system or party to another. These transmissions can take the form of a file or “feed” sent from one system to another, or direct access from a UI, a service, etc… In addition, key object properties such as *Sources* and *IsSourcedBy* are necessary to show which meta-data-artifacts serve as the source and target for various transmissions. As with other classes, transmissions have attributes and the provenance properties can be assigned at any level of detail. In some cases it may be easier for a developer to state that the contents of the Customer table and a SalesPerson table source a particular data feed that has that information combined, rather than stating the dependency at the column level. In several cases, knowing just the involved tables can provide enough information to perform impact analysis of changes, data flow analysis, etc. Obviously, having column level detail makes for a more robust knowledge base, however that can be achieved over time as the organization gets the benefits of the knowledge base much sooner.

Governance Entities

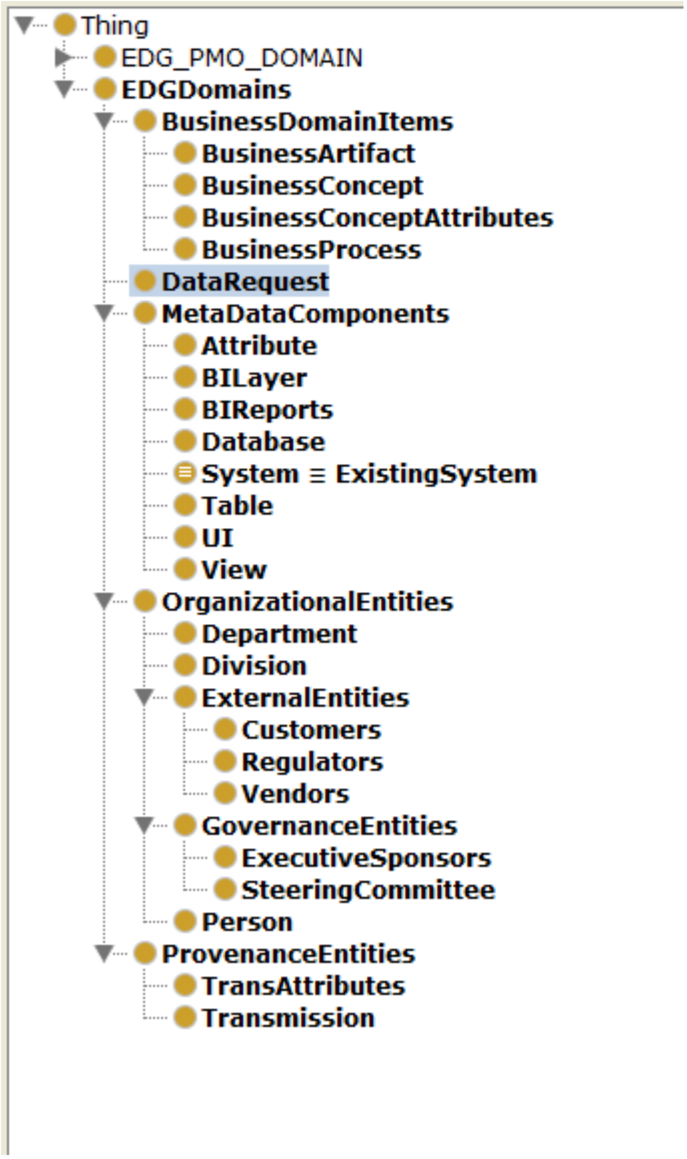
These entities can be arbitrary groupings or committees that oversee the lifespan of various systems and business processes. This class differs from the organizational entities in that it is more variable and dynamic. Governance entities can appear for given projects, and have certain object properties such as "must approve" or "must be notified" that link individuals to data. Organizational entities however tend to map more closely with the organization as a whole to maintain a consistency throughout the ontology. Additionally, depending on particular organizational policy, complex roles and relationships can be designed to link organizational entities to corporate data.

Business Domain Concepts

The business concept class is to capture business definitions and processes that exist in the organization and drive the systems and the data they produce. Here, we will find that this class takes full advantage of semantic web by allowing various groups throughout the organization to specify the concepts that they use and or own and input as much data as possible. The eventual goal is to not only align business concepts and processes to the systems and data that make them up, but to also align various concepts across the firm that may refer to the same thing. Business activity such as mergers serves as a very common use case for this situation. Here, after a merger a business concept such as *Account* may mean one thing, however map to the Accounts system of each of the merging companies, and possible another system with a combined population. Once again, the strength of the semantic web technologies is in the flexibility that we have when populating data. In most cases conceptual definitions can lead to much discussion and debate. Here, we can start populating the knowledge base with less "debatable" information such as organizational structure and meta-data and realize benefits quickly. Once the physical aspects of the data landscape are modeled, then the conceptual or semantic entities can be focused on. This is a common pattern in the area of dataspaces where decisions on semantic definitions are postponed while the "known" world is modeled and generating a return on the investment.

Request Entities

These classes are indented to capture data from a requesting point of view. Should the organization want to capture meta data requirements at their inception, they would need to have a facility intake and rout data requests to the appropriate data owners. These requests would then go through an approval process and eventually result in a transmission. When an organization chooses to adopt such detail they will then be able to track data movements throughout its total lifespan. Facilitating data requests not only hinges on data governance policies adopted by the firm, but as the framework is adopted by the organization, these entities and properties can serve as a means to ensure clean intake channel for a robust governance process.

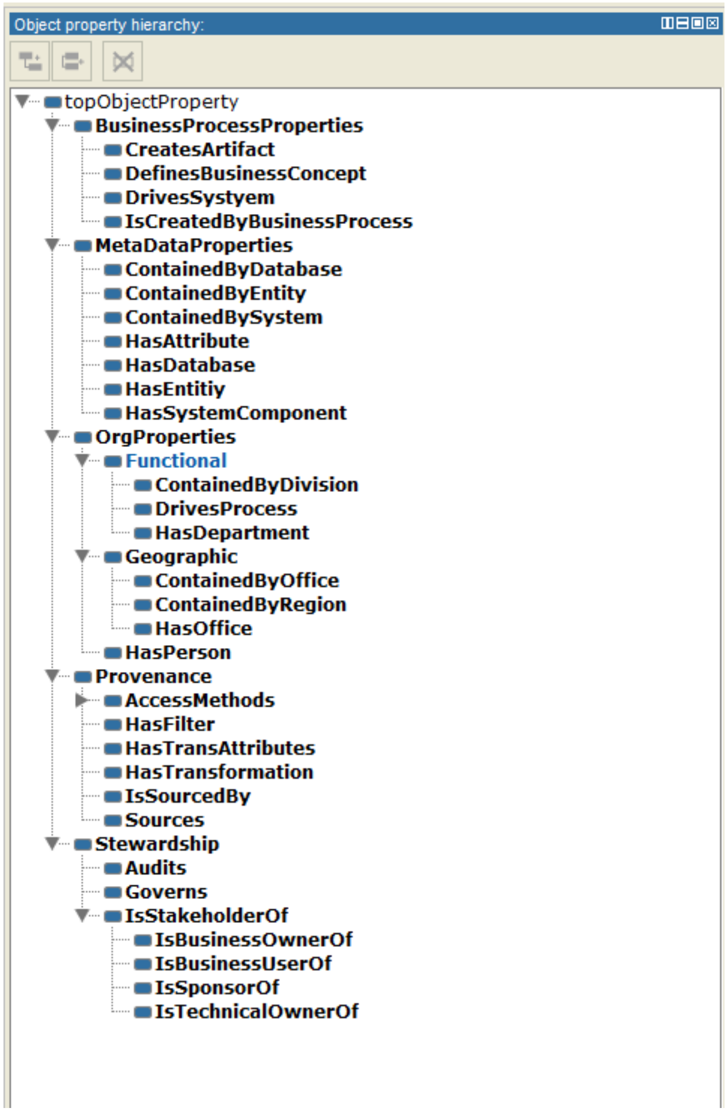


- Thing
  - EDG_PMO_DOMAIN
  - EDGDomains
    - BusinessDomainItems
      - BusinessArtifact
      - BusinessConcept
      - BusinessConceptAttributes
      - BusinessProcess
    - DataRequest
    - MetaDataComponents
      - Attribute
      - BILayer
      - BIReports
      - Database
      - System ≡ ExistingSystem
      - Table
      - UI
      - View
    - OrganizationalEntities
      - Department
      - Division
      - ExternalEntities
        - Customers
        - Regulators
        - Vendors
      - GovernanceEntities
        - ExecutiveSponsors
        - SteeringCommittee
      - Person
    - ProvenanceEntities
      - TransAttributes
      - Transmission

Figure 16 Protégé Class View

3.4.5 Object Properties

In addition to the classes, which we have included in our ontology, there are also a several different object properties that are used to model the relationships that bind the organization and data in general. Some of the properties are used for the purposes of modeling a containment relationship within the classes. This is because the to default relationship for OWL classes is one of inheritance. However, there are others, primarily in the areas of stewardship and governance that are used to capture the flow of data throughout the firm and all involved parties.



- topObjectProperty
  - BusinessProcessProperties
    - CreatesArtifact
    - DefinesBusinessConcept
    - DrivesSystyem
    - IsCreatedByBusinessProcess
  - MetaDataProperties
    - ContainedByDatabase
    - ContainedByEntity
    - ContainedBySystem
    - HasAttribute
    - HasDatabase
    - HasEntitiy
    - HasSystemComponent
  - OrgProperties
    - Functional
      - ContainedByDivision
      - DrivesProcess
      - HasDepartment
    - Geographic
      - ContainedByOffice
      - ContainedByRegion
      - HasOffice
    - HasPerson
  - Provenance
    - AccessMethods
    - HasFilter
    - HasTransAttributes
    - HasTransformation
    - IsSourcedBy
    - Sources
  - Stewardship
    - Audits
    - Governs
    - IsStakeholderOf
      - IsBusinessOwnerOf
      - IsBusinessUserOf
      - IsSponsorOf
      - IsTechnicalOwnerOf

Figure 17 Protégé Object Properties

The following Object Property groups are defined in the core ontology:

Stewardship

The primary purpose of these properties is to model the relationship between organizational entities and the meta-data entities. The properties were from established data governance concepts [4, 12]. Here, we identify which people or teams are designated as stakeholders, owners, stewards, etc. This is very useful when doing impact analysis for changes and approvals are needed from various parties who are involved with the data and a crucial underpinning to any data governance initiative. Additionally, when issues occur that need forensic analysis, identifying the individuals or areas involved will be as straightforward as querying the objects of this property with a given meta-data source as the subject. On a day to day basis, having this knowledge documented will allow for better decision making as an ongoing process as the parties involved with the data will be publically known to anyone browsing the ontology.

Provenance

Provenance properties are essential in modeling the lifespan of data [10]. In addition to properties that describe containment, the property *Sources* and its inverse, *IsSourcedBy,* are essential in modeling data dependencies. Here, a system or database can be shown as source or a dependent system in relation to another. When used in conjunction with the *Transmissions* class, a particular data feed that many systems are dependent upon can be easily modeled in a hub and spoke fashion. As with other classes, this property can be applied at lower levels of detail. When used at the columnar level, the most detail, including transformations can be conveyed. In other cases, however, having this information at the tabular level, or even database level can provide enough information for managing data.

Business Process

Business process object properties envelop most of the classes in the ontology. They are used describe the relationships between business processes and the artifacts that they produce, as well as their relationships to the systems that are used to support those processes, and the functional areas within the organization that are responsible for those processes [88]. Additionally, these properties are deemed reflexive in that they can apply to other business process. It is often quite common for one business process to depend on the artifact of another. Having a keen understanding of which business process rely on which data help in ensuring that data governance is not only an "IT" issue rather it cuts across organizational structure, requiring maximum participation in order for it to be successful.

Meta-Data and Organizational Structure

Organizational properties are primarily for establishing a tree-based hierarchy that usually reflects most corporate organizational structures, both from a functional and geographic point of view. Similarly, Meta-Data properties describe the parent child relationship that most systems exhibit. This is primarily driven by relational database concepts such as a database containing tables and views, which further contain columns.

3.5 Operational Guidelines – Setting up the Solution

We have seen in the prior sections the primary components of the solution framework. This next section describes the steps that need to be taken and decisions that need to be made in order to configure those components and operationalize the approach.

3.5.1 Designing the Ontology

The fist step in adopting the solution is deciding on how the core ontology should be constructed so it best models the organization to the level of detail required. As we have seen in the prior section, the nature of OWL, and RDF, lend themselves to flexible adoption. This means that not everything needs to be stated up front, however a general strategy needs to be adopted in order drive how the ontology will be used and interact with individuals. In the ontology that was created for this research, classes were derived based on fundamental aspects of data governance that center on aligning the business with technology. As the basis of this research is to improve data governance, the reader should start with identifying classes that capture how data assets and the organization interact. Referring to Figure 16, the classes that will accomplish this are:

- Organizational Entities: These classes will capture the divisions, groups, sub-groups, departments, and even people if that level of detail is desired. The goal is to model the organization at a level of detail that is both informative and accurate, but also maintainable. The implementer of this solution needs to decide if it makes sense to mimic the organizational structure as per other enterprise systems, such as finance and HR, or to create an entirely new structure that pertains to this initiative.

- Meta Data Entities: Perhaps the most self-explanatory case. Here a class needs to be established that actually represents the data assets themselves. As with the prior group of data, the level of detail needs to be established that is both informative and sustainable. It is recommended, however that to truly identify potential impacts, and have a clear awareness of the flow of data elements, that field or columnar level detail is embraced for this subject area.

- Business Domain Entities: While it is critical to show how data assets and business initiatives interrelate, from a tactical perspective, the most important of these classes are the ones that represent existing business processes. This will further allow for connecting how the business's actions rely upon specific data assets.

In addition to the classes, it is also critical to establish the appropriate object properties to relate the classes and their subsequent individuals together. Establishing the following object properties are critical:

- Organizational Properties: These object properties, such as "Contains" allow for hierarchical corporate structures. Other properties such as "Works With", or "Supports" can be used to highlight matrix style relationships.
- Governance / Stewardship Properties: These properties are critical in relating data assets to the appropriate people or groups within the organization. Here, it is important to use relationships that properly reflect the strategy for governance, such as data "Owners" and "Stewards" and "Sponsors". Additionally, terminology can be used to identify parties that are Responsible, Accountable, Consulted, or Informed, (otherwise known as RACI) regarding the data assets.
- Provenance Properties: Identifying the properties that model the lineage of data assets is critical in showing the supply chain of data. Concepts such as

"sources" and "is sourced by" will highlight dependency. It is also important to ensure that these properties are notated as transitive so they can depict a property chain[33].

3.5.2 Populating Core Classes

Once the ontology has been established, it needs to be populated with the appropriate data relative to the classes and properties that have been chosen. Here, the high volume classes are the organizational structures and the data assets. Depending on the size of the organization, number of systems/assets, and level of detail that is chosen, this task can range from several hours to weeks.

Generally, obtaining the class information is a straightforward, albeit sometimes tedious task. The act of relating that information via the object properties is where there needs to be research, agreement, and some degree of automation. Organizational information can be gotten from core systems such as the General Ledger, or HR Information Systems. Systems, databases, tables, queries, and fields can also be gotten in some automated fashion as well, such as querying systems catalogs (see chapter 4). Some organizations may also possess an application inventory system, which may have some ownership information as well. Other classes, such as business processes may need to be gathered manually as they are somewhat abstract to be contained in an inventory system.

Once the classes have been loaded with the appropriate individuals. They need to be related and formed into "triples" via object properties. Here, once the due diligence has been performed which identify the appropriate relationships between various individuals, there needs to be a means for generating this information and loading it. For the ontology used for this research, Microsoft Excel was used, in conjunction with text handling formulae to create the appropriate triples.

3.5.3 Decide on Level of Centralization

While this is more of an organizational consideration, it also has implications on how this solution will be implemented. Ideally, once the core classes have been created, distributing the ownership of this initiative and putting the onus on local groups of individuals to itemize their data assets, systems, business processes, and sponsors, is where the strength of linked data comes to play. It is however sometimes easier to start small with one or two areas in a highly centralized manner prior to spreading out to other groups. The core ontology, however is best kept centralized so it remains consistent for other groups to reference. It is recommended that a rollout plan be established after key setup is done to allow other groups to adopt the platform and add their information to it. As with most crowd-sourced solutions, the value of this approach will further increase, as more groups come on line.

3.5.4 Identify Strategy for Keeping the Model Current

When abstractions are used to model the enterprise, it is important that there be a strategy to keep them current, lest credibility is lost when it is no longer an accurate representation. This can be seen when documentation falls out of date, relative to the code it is describing. Also, when data modelers use artifacts such as ER diagrams to model a relational database, they must be kept up-to-date in order to remain useful to the broader population. In this case, it is critical that the ontologies accurately represent the organization. In most cases, this will require a degree of automation in place to appropriately update it when there are changes to the organizational structure, or new data assets are created, etc.

Regardless of the approach taken to maintaining the models, the level of centralization adopted also pertains to this as well. If a highly de-centralized approach is taken, then the respective areas must ensure that their "local" models are as accurate as possible. With everyone embracing their own model maintenance, referring to other ontologies can be done with a higher confidence level. Furthermore, with a de-centralized approach, the level of modeling for a given group such as a department would be less, as data assets would be contained to what that group owns and or uses.

3.5.5 Identify How Data Movement will be Captured

Data movement is perhaps one of the more abstract elements to capture with this framework. Once again, depending on the level of detail that the organization wants to embrace, there can be several ways to approach it. In one case, it can be as simple as having one data asset, such as a table, or a field, or a view *source* another data asset. Here, the implementer can choose to be at a low level of detail and specify fields, or remain somewhat abstract and identify that a given table may source another system. This is the tradeoff that must be made, as the finer detail provides more granularity, it also has a higher maintenance effort. Often times, knowing that a given table may be involved in a particular interface to another system, may be enough visibility for an organization. As with graph databases however, the high degree of flexibility can allow groups to choose a level that suits them.

Another decision point is also creating a concept of *Transmission* to abstract data movement. If a particular set of data is re-used by many groups within an organization, such as an employee roster, it could reduce maintenance to closely model that data set as a *Transmission* and have it sourcing multiple systems. This abstraction can be quite helpful in organizations that have several different methods of moving data, such as file "feeds" in which a flat file is sent to another system for ingestion, or via an API, or exposed web services, or perhaps a database view. The modes of transfer are quite different, however stewardship and dependency information are the same.

3.5.6 Identify Querying and Visualization Methods

Once the major aspects of the solution are put in place, the organization must decide how this data is to be used. In chapter 4, we will review tools and methods for querying data from the graph. Regardless of implementation however identifying operational use cases, such as daily usage reports, or exception reports, or data dependency graphs is critical to ensure the solution is providing value to the day-to-day operations of data governance.

3.6 Data Governance Items Addressed

3.6.1 Introduction

While the immediate goal of this paper is to provide a method for increasing data awareness in organizations, the larger goal however is to use that increased awareness to improve data governance. In much of the literature that exists regarding data governance, key success criteria is defined as addressing the issues that exist because of a lack of it completely, or improper execution. Sarsfield best sums up the success criteria for data governance across several key areas which will be described in the following sections [3]. We will now describe how this solution assists in these areas of data governance.

3.6.2 Fixing Data Anomalies

Data anomalies can generally be categorized as “bad data”. Examples of this is missing or incomplete data, illegible data, the wrong data types being put into fields, data that is in the wrong format for SSN or phone number. The cause of the anomalies can be anywhere along the data supply chain. If the data in question is sourced from another system via an interface, then, the root cause could lie in the transmission process. Additionally, if the source of the data is directly coming from a system, the user interface may be the cause. An example of this would be an Applicant Tracking System that has a form for perspective job applicants to fill out, but has long text fields for educational information instead of discrete lists of universities, degrees, and majors. The long text field can result in people entering "MBA" in some cases, or "Masters of Business Administration", or "Masters of Bus. Admin."

While the act of correcting the data and the related process, is outside of the scope of this research, this solution allows the data governance committee, or other concerned parties to identify processes, transmissions, and data lineage in order to help ascertain the effort and touch points associated with the fix.

3.6.3 Defining a Repeatable Process

As we have seen in Chapter 2, defining a repeatable process is a key component of many of the data governance maturity models. The concept of a repeatable process ensures that data governance activities are not "one off" exercises. The act of defining methods to keep this framework and the representative models current with the actual state of the organization contributes to a repeatable process. Additionally, as processes may be put in place to clean data, or that continually monitor data for quality and make updates as needed, they too can be part of the overall framework as they become part of the data supply chain, as they are modifying data. Having these processes captured as any other assists in documenting their role in the data lifecycle.

3.6.4 Handling Change

One of the main aspects of good data governance is to be able to adapt with the changing business climate. Often times, applications, interfaces, external SaaS solutions change behavior, processes, etc. resulting in a lasting data change. Unfortunately, the nature of data architecture, with all its dependencies is less nimble and more susceptible to risk of change. Here, the framework and increased data awareness allows the organization to be much more proactive in determining the impact of change. Even if the organization has embraced capturing less detail in this solution, having a record at higher levels of detail will provide some insight into impacts. Additional due diligence can then proceed in a more localized fashion.

With regular reporting and visualizations will also come a heightened awareness of the data landscape outside of the traditional central team, e.g. the Data Governance Committee. Having more individuals involved in data awareness will provide more knowledge when change comes to the organization.

3.6.5 Coordinating Efforts with the Business

The nature of data governance is to control data in the organization throughout its lifecycle as with any other corporate asset. The result is a combined effort with the data stewards (typically IT) and the data owners (typically the business). In the past, data initiatives were often thought of as technology endeavors, however as data tends to be the "footprints" of many processes, it encapsulates a great deal of business knowledge which is necessary to help manage it. This is especially true when looking at data over a long period of time, as processes and business conditions may have changed over that span, with no other trace to mark that change except for certain data values.

Capturing abstract concepts in the ontology, such as Business Processes, or Business Concepts and a providing a means to relate them to actual technology assets, organizational entities, and projects, greatly aids in coordinating and aligning the business with data. Often times a business area will be unaware of all of the processes, systems, and data under their span of control. Increasing this awareness will help foster alignment across functions.

3.6.6 Fostering Data Ownership

Data ownership is a major aspect of governing it. All too often, data is "assumed" to be owned by someone. This lack of responsibility often causes the technical owners of the systems to become the de-facto owners of the data itself. The result is a party that doesn't fully see the context in which that data is used in an operational context. This lack of ownership and accountability can lead to irresponsible use of the data and decision making that doesn't always take all the stakeholders into consideration.

This solution, through ontology puts a great deal of emphasis on data ownership via organizational classes and object properties that define relationships to data assets. Additionally, the ontology allows for a clear delineation between stewards, stakeholders, and owners so all parties involved are aware of how they relate to the data assets. The sense of ownership can also be reinforced through various visualization tools, which can visually identify individuals and how they relate to various data and technology assets.

3.7 Solution Benefits

As stated in chapter 1, there are several scenarios where such increased data visibility throughout the organization would be beneficial:

- Proactive Impact Analysis: Being able to identify the affected systems, processes, and parties before changing a piece of data or the process that creates that.

- Reactive Impact Analysis: Knowledge of the full lineage and involved and responsible parties around data elements will aid in addressing and rectifying any unexpected issues that come up around that data.

- Forensic Analysis: Management or regulators may want to have visibility into where data is flowing from a system standpoint, but also an organizational and geographical standpoint.

- Identifying business definition consistency issues: Having a clear way to see what data is being created and which systems are the systems of record for those data elements will assist consuming application that hey are pulling the most authoritative data available for that subject. Also, having the parties associated with that element f data will also provide additional information as to the business definitions behind that data.

- Managing IT Costs: Such data visibility will allow management to see which areas of the firm are maintaining redundant data stores, or outdated interfaces to multiple systems.

- Risk mitigation: Having a means to tag particular data elements as sensitive and monitor their data flows as well as requests for that data provide additional checks to mitigate risk of data leakage to the wrong parties.

- Opportunity Identification: Having visibility to enterprise data and its movement throughout the organization will allow for groups to see what data is available throughout the firm as well as management to see which groups are using that information in an effective manner and if so, can it be an opportunity for improved systems or business growth.

- Business layer abstraction: Having meta-data appropriately modeled provides the foundation for creating business layer abstractions on top of those physical representations that can be shared across the organization. The business intelligence ecosystem has many tools, such as Business Objects, or Oracle Business Intelligence[100], or Microstrategy [101]. The common theme across those is a means to abstract the physical aspects of data, such as their tables, columns, foreign keys with a "semantic" or "logical" abstraction layer so business users can extract what they need without having to know about the details of how the data is stored or organized. Generally this process is iterative and requires a good degree of human interaction to define the appropriate definitions of certain items. Semantic technologies allows for this process to occur in such a fashion as individuals can contribute in a collaborative and ongoing "pay as you go" basis. [102]

In the above scenarios there are several types of consequences that the organization faces for not having the appropriate awareness of corporate data:

- A scenario can be created where data quality is compromised leading to further impacts throughout the firm.

- If already compromised, the organization's ability to quickly contain the impact can be reduced.

- The organization can face regulatory risk by not being able to quickly produce a view of where sensitive data throughout the firm is traveling.

- Redundant data stores and obsolete interfaces can increase IT costs while also increasing the likelihood of data leakage throughout the organization.

- Potentially missed opportunities to expand corporate knowledge based on current data usage terms or not having a foundation for business abstraction layers.

3.8 Conclusion

We have seen that there are consequences for large organizations to not have a solid understanding of the data it has, where it resides, and who is accessing it, how, and for what reason. Their ability to effectively govern their data becomes quite challenging, with the result of poor governance impacting overall data quality and increasing risk throughout the organization. While there exist tools, packages, and processes to address this situation, getting the participation of the key data stewards and owners is paramount to building this vision. Given the current structure of today's large multi national companies, cataloging their data tends to be a second or third tier responsibility of many of these data stewards. Additionally, as data management projects tend to involve many parties and require costly packages, IT management is often reluctant to fund such initiatives.

It is the goal of this research to illustrate that semantic web and linked open data can provide a flexible, incremental, low cost, and low barrier to entry manner in which to increase enterprise data awareness and improve data governance while providing an extensible foundation for further initiatives.

Chapter 4

Implementation Overview

4.1 Introduction

The goal of this research is to illustrate a flexible, incremental, and economical means to gaining awareness of corporate information assets. To that end, there are several different approaches and tools that can be leveraged to implement the proposed solution components as described in chapter 3. In this chapter, we will go through some of the implementation concerns that would occur across multiple tools. Like most modeling tools, in order to fully realize the benefits of semantic web as a framework for modeling enterprise meta-data it is imperative that the graph structures accurately reflect the meta-data throughout the organization. First, there must be consistency with the model and the physical data that it represents. Second, there needs to be a way to ensure all of the component ontologies are consistent with each other in terms of versioning and vocabulary.

To achieve this, we can approach it from both a proactive and exception based point of view in which consistency checks can be applied on a scheduled basis, as well as providing the means to collect information from the users directly. The nature of this solution affords the implementers flexibility in how much they want to embrace over time. Much of the integrity checking can be conducted manually, or with “low tech” means, such as scripts and or office tools such as MS Excel, however there are means to construct fully automated and aware approaches to this solution. As stated in the prior paragraph, exception based approaches can also be taken in which manual updates to the graph structures are manual, but automated “audits” are run to check the integrity of the solution. The approach will be dictated by the size and complexity of the organization, as well as their aggressiveness in pursuing the solution. This chapter will review the implementation highlights of the various solution components, as well as describing key features of the tools and utilities that are available to construct it.

4.2 Solution Considerations

In this section, we will look at elements to keep in mind when implementing this solution in a production environment. While there are some elements of the framework that need to be kept timely and accurate, the nature of linked data and the solution architecture lend themselves to flexibility, so organizations can incrementally add to the system as it suits them.

4.2.1 Graph Maintenance – Accuracy

As the enterprise changes during the normal course of business, so too must the model that represents it. While there are some types of organizational information that can be loaded in an automated fashion, such as physical data components like tables and columns, and organizational information such as people and departments, other concepts such as stewardship and provenance may only have to be updated on an as needed basis, or require human interaction to complete. An example of such a change is if new individuals become data stewards or if there are changes to the organizational structure. To this end, two approaches to keeping the graph structures current are proposed.

The first method is via an automated process, such as an agent, “crawler”, or scheduled batch processes. This approach will allow for the automatic updating of the various graph structures with key, organizational information. While any of the classes within the ontology can be updated, the most common areas would be the ones that are highly structured and standardized, such as core data structures and organizational information. In some cases, organizations have rigorous project, portfolio, and process governance approaches, so that information may be readily available as well. The following example describes how this process would be employed with regard to corporate meta-data describing data assets:

The crawler is an agent that is intended to run on a scheduled basis and has the task of ensuring the meta data components specified in the ontology are appropriately reflecting the RDBMS objects that exist. This would entail iterating through the corporate ontology to identify the physical databases, which must be traversed. Then the database catalogs for those databases would be queried and compared with the appropriate ontologies. Databases, tables, views, and columns would be checked against the ontology and an alert would be triggered should there be an inconsistencies. For database views, the crawler will comb through the supporting SQL and pull the appropriate columns from the "select" statement, and can identify the supporting tables in the "where" clause. For a more automated approach, the crawler can make the changes to synchronize the two and log the actions appropriately. In addition to serving as an automated reconciliation, the crawler can also be effectively used to "seed" the meta-data portion of the component ontologies for a given area. Some database schemas, particularly those of ERP systems can have a large number of tables, hence an automated way to populate the system is an efficient way to jumpstart corporate adoption. This approach is similar to how database designers keep their models in sync with reality by "reverse engineering" based on the database structure.

The same approach that is used with corporate meta-data can be taken with any stored information, such as organizational and departmental information. Corporate actions such as divestitures and restructurings alter the shape of the organization and workers' responsibilities. In order to be a tool to foster data governance, the organization must be accurately reflected. Here, the process will most likely query the organization's Human Resource Information System (HRIS) in order to get that information and verify it is accurate or update accordingly.

The second method for keeping core data accurate is via a user interface (UI) for manual intervention. There are entities that will require human intervention to appropriately model, such as governance committees. Often, this type of information is generally not stored in any corporate system, rather informally known across teams. In this case, the graph itself could potentially be the system of record for this information. The UI will allow for the change to be made in one place in a standardized format. Furthermore, the ontology also contains classes to align business concepts and processes with underlying data. Here, abstract items such as "account", "customer", "product development process", "employee hiring and on-boarding" can be stored and linked to the underlying systems that either represent the concepts themselves, or are involved in the processes. Often, populating the individuals within these entities will require collaboration with the business and their technology counterparts. In essence, the framework becomes the system of record for these organizational concepts.

As stated before, the model is only useful if it accurately reflects reality. In an ideal state, all changes will originate via the UI and there will be very tight integration with the change management processes in the organization. However, in keeping with the concept of data spaces [28] an incremental approach will be the most realistic to for organizations to undertake. Once the meta-data is populated and maintained, a given department can over time enrich their component ontology with more data about stakeholders, business concepts, data feeds, etc.

4.2.2 Graph to Graph – Consistency

When implementing the graph throughout the organization there are two primary ways in which the model can be persisted. The first is as a central approach where all corporate data as well as departmental system information would reside in one store such as a serialized file, or a triple-store. This approach has the benefit of enhanced performance for querying as there is no federation to take place, as well as more control on the integrity of the content. Individual departments that would contribute their information would need to be able to update their section of the graph, which can be done independently and integrated at a scheduled time, such as in the evening. This approach would require additional process to perform the integration with the corporate ontology. As the data is centrally located, business continuity is a more straightforward task, as there is only one location for which there needs to be redundancy. The detraction from the centralized approach is that groups and teams must coordinate with a central body to ensure their data is being integrated properly and at timely intervals. While this provides some consistency, it also impinges on the flexibility provided by linked data.

The other approach is to have individual teams publish only the information for which they have, such as their own meta-data about underlying systems and integrate by linking dynamically to the other graphs throughout the enterprise. There would still need to be an enterprise graph that would contain central information such as business concepts and organizational information that will serve as the common thread for all other team's content to link to. The benefit to a decentralized approach is that groups can proceed to publish information about their area rapidly and at a low cost. Teams would have to have a disciplined approach to maintaining their data as it get linked to. Additionally, as there is no central store for housing links between groups, ground rules and standards need to be in place to ensure consistent usage and expansion of the vocabulary. For example, if department A sends a feed to department B, it is essential, that one party "own" the URI of that transmission entity (typically the source area). The target area can then reference the appropriate URI. The decentralized approach also allows the graph to grow in a more organic way throughout the enterprise, encouraging more participation. Version control is essential as well to ensure that all groups are working with a vocabulary with the same classes, constraints, and axioms. As one of the principles of linked data it is important that data contributors seek existing URI's or vocabularies first, before creating their own [75]. Unlike the internet, being in the "closed" environment within the organization provides an advantage in that a consistent vocabulary can be used across all parties, reducing the effort required to identify entities a semantically equivalent to one another. This consistency of vocabulary is very important to fostering rapid adoptions throughout the enterprise.

As we have seen in prior sections, the strength of the semantic web is that enables a decentralized approach at sharing knowledge. This is seen in the ability to publish an ontology as well as reference any URI that is accessible and integrated within your ontology. Furthermore, the rich expressiveness of OWL allow for statements that align concepts that may be named differently. In this research, it is assumed that one vocabulary will be shared across the enterprise. We make this assumption as this is an organization and standards can be applied in essence creating a "closed world". Additionally in order to effectively look at the flow of data and query the information in a fast paced business setting, a consistent, standardized, and fully agreed upon vocabulary needs to be established.

While the vocabularies will remain consistent, the individuals within the ontologies will vary widely, reflecting the relational data base landscape of each group. The semantic web allows each team to have the ability to publish and maintain their metadata on a scheduled basis that suits them best. In addition to locally specific data, in order to fully bring all of this disparate data together, we need to represent the entities that are common across the enterprise, such as governance data and the organizational information. Therefore, having a corporate ontology that is populated with this centralized data, and serving as an "integration registry" will maintain the consistency that is required to effectively capture data stewardship and data flows. Given the decentralized nature of publishing data, along with a central store of common data, it is essential that the crawler also reconcile the component ontologies with the corporate ontologies to ensure proper versioning, ensuring that data exchanges ("transmissions") are appropriately balanced from the source and target system's point of view. Similar to its function when reconciling databases the crawler can either alert the administrator to inconsistencies in the ontology or if enough information is present to make a decision, take corrective action

4.2.3 Data Lineage

In addition to modeling the enterprise meta-data, key provenance information such as data requests and data consumption need to be captured as well to effectively capture dependencies, lineage, and the overall flow of data throughout the organization. Much of this kind of information requires human interaction in order to initiate the data request and subsequent approval processes that must take place in order to effectively govern metadata.

As such, it is an area that will require appropriate data governance processes. As is the nature of ontologies, data can be updated in a distributed and decentralized fashion; it is imperative to provide a means to evaluate the various components of the graph to ensure consistency.

Provenance information and their associated data requests are perhaps one of the more dynamic features of the UI. As organizations do business in an increasingly global environment, their organizational structures are getting more diverse. It is very important for teams, departments, groups, etc. be able to identify data that is of use to their line of business and request access to it in a timely fashion. Just as important however is the ability to have an approval process around that request to ensure the correct data is flowing to the correct recipient, who has the authorization to see such data. Additionally, to ensure transparency into the lineage of data, the request and it's associated workflow need to be stored and along with it's corresponding description of that data transmission.

Generally, when an individual requests data they often are only familiar with the overall system or group that they need data from. Hence, the nature of the process is one that is iterative and requires communication across the teams. The UI presents the requestor with the various systems and hierarchies of meta-data available for them to browse. In most cases, a request can be initiated that pertains to an abstract business concept, which, as per the object properties stated in the ontology, in turn drives systems, which in turn have technical owners to be contacted. If the requestor however is more knowledgeable of the area from which they are seeking data from, the request can specify lower levels of detail, thereby reducing the time to finalize the details of the data transmission.

Concurrent to finalizing the transmission, an approval process can start. The UI will provide a means to view all pending data requests that require approval before development can start on them. Once development has been completed and the data transmission individual is associated with the appropriate sources of data (e.g. tables, columns), all the parties in the process are identified and associated to their role. This combination of tracking request and approval is essential in capturing the appropriate lineage of data while also enforcing a degree of governance and transparency into the movement.

4.3 Implementation Examples

As described in the prior section, while there are elements that are central to this solution that transcend the tools, there are numerous approaches that can be taken to implementation which allow for an organization to adopt as much or as little of the solution as possible. The following points illustrate various tools that can be employed in this solution.

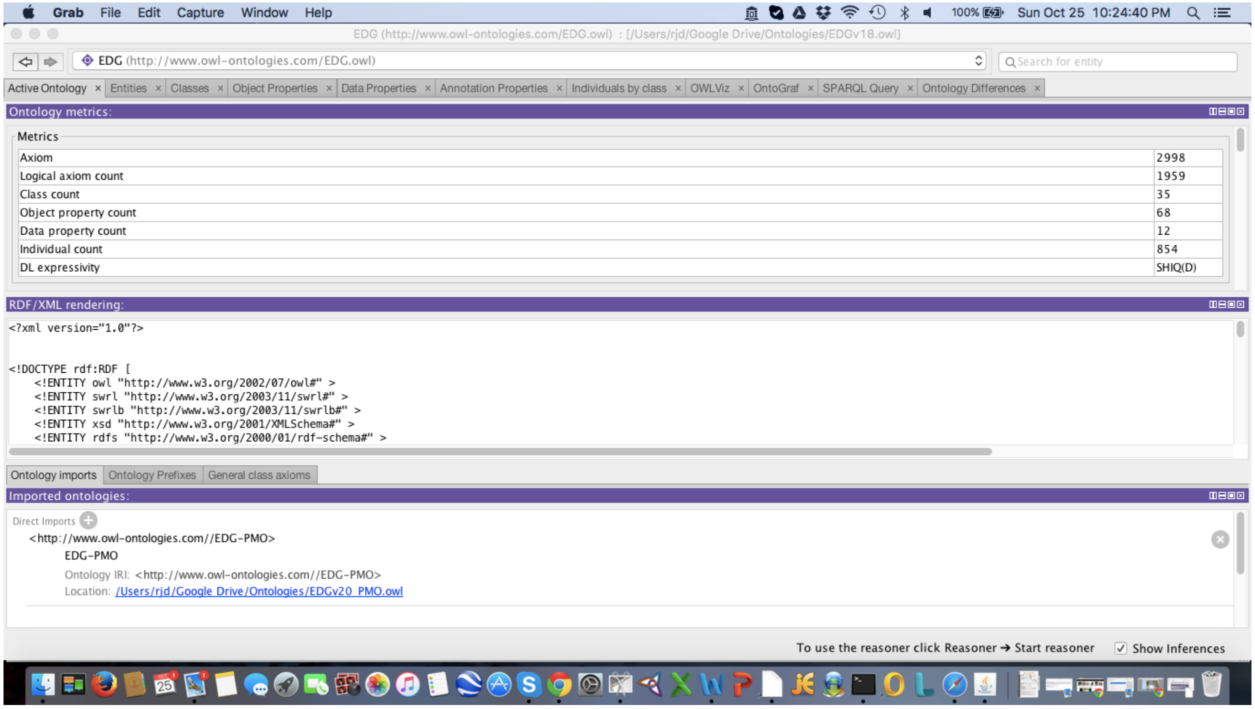
4.3.1 Visualization and Modeling Methods

One of the primary aspects of this solution is to be able to visualize the landscape of data assets within an organization. Additionally, there must also be a means to not only view the graph of data, but also be able to manipulate the ontology that serves as the basis for the classes and their associated relationships.

To craft the ontology for this research, the tool Protégé [103] was used for knowledge and domain modelin. Protégé has both a desktop version as well as a hosted solution to allow for easier sharing, collaborating, and viewing ontologies across disparate groups. For this research, Protégé 5 on Mac OS was used. It is fully supportive of the latest OWL 2 Web

Ontology Language, as well as many serializations such as RDF/XML, Turtle, and OWL/XML.

The main components of Protégé are the browsers that allow you to add and manipulate both your ontology classes and object properties. Here is the primary screen that shows the active ontology that is loaded into the application. While the screen is highly configurable, it shows key information such as metrics around the size of the ontology, a rendering of the ontology, in this case, RDF/XML, and information regarding any ontologies or data sets that are being imported or referenced by this ontology.



Metrics

Axiom	2998
Logical axiom count	1959
Class count	35
Object property count	68
Data property count	12
Individual count	854
DL expressivity	SHIQ(D)

```
<?xml version="1.0"?>


<!DOCTYPE rdf:RDF [
    <!ENTITY owl "http://www.w3.org/2002/07/owl#" >
    <!ENTITY swrl "http://www.w3.org/2003/11/swrl#" >
    <!ENTITY swrlb "http://www.w3.org/2003/11/swrlb#" >
    <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
    <!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#" >
```

<http://www.owl-ontologies.com//EDG-PMO>

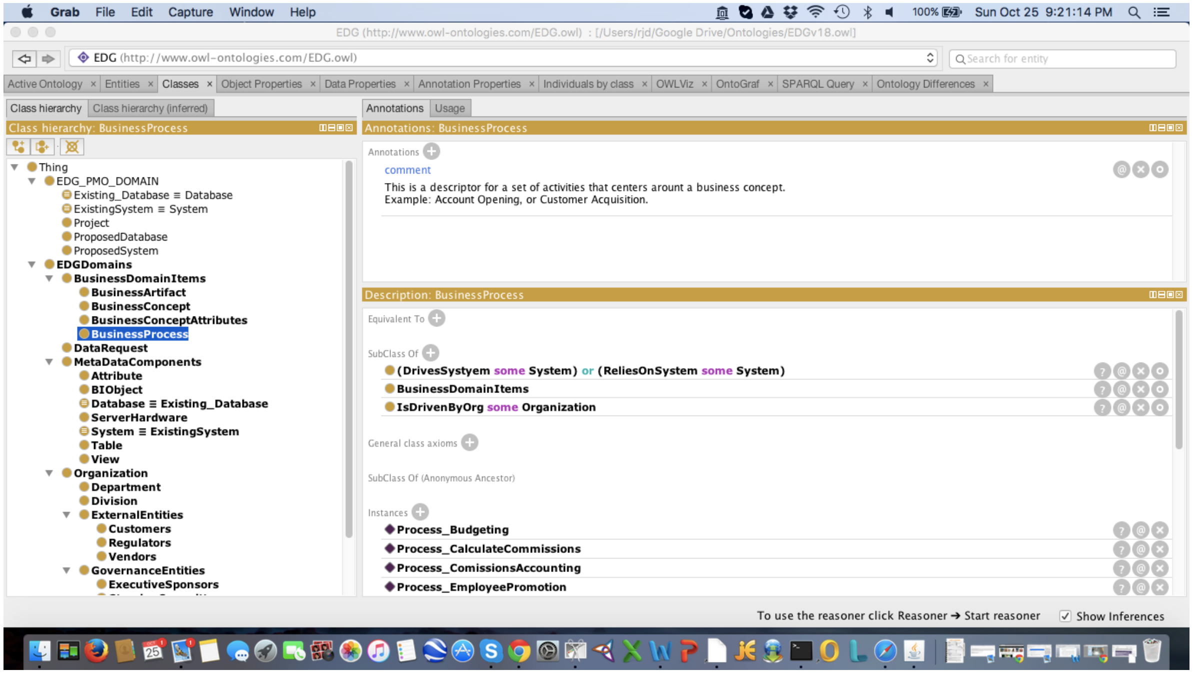
EDG-PMO

Ontology IRI: <http://www.owl-ontologies.com//EDG-PMO>

Location: /Users/rjd/Google Drive/Ontologies/EDGv20 PMO.owl

Figure 18 Protégé Active Ontology Screen

The following example shows the primary viewer for ontology classes:


Grab File Edit Capture Window Help
100% Sun Oct 25 9:21:14 PM
EDG (http://www.owl-ontologies.com/EDG.owl) : [/Users/rjd/Google Drive/Ontologies/EDGv18.owl]
EDG (http://www.owl-ontologies.com/EDG.owl)
Search for entity
Active Ontology × Entities × Classes × Object Properties × Data Properties × Annotation Properties × Individuals by class × OWLViz × OntoGraf × SPARQL Query × Ontology Differences ×
Class hierarchy
Class hierarchy (inferred)
Annotations
Usage
To use the reasoner click Reasoner → Start reasoner
☑ Show Inferences

Class hierarchy: BusinessProcess

```
▼ Thing
  ▼ EDG_PMO_DOMAIN
      Existing_Database ≡ Database
      ExistingSystem ≡ System
      Project
      ProposedDatabase
      ProposedSystem
  ▼ EDGDomains
    ▼ BusinessDomainItems
        BusinessArtifact
        BusinessConcept
        BusinessConceptAttributes
        BusinessProcess
      DataRequest
    ▼ MetaDataComponents
        Attribute
        BIObject
        Database ≡ Existing_Database
        ServerHardware
        System ≡ ExistingSystem
        Table
        View
    ▼ Organization
        Department
        Division
      ▼ ExternalEntities
          Customers
          Regulators
          Vendors
      ▼ GovernanceEntities
          ExecutiveSponsors
```

Annotations: BusinessProcess

Annotations

comment

This is a descriptor for a set of activities that centers arount a business concept.
Example: Account Opening, or Customer Acquisition.

Description: BusinessProcess

Equivalent To

SubClass Of

- (DrivesSystyem some System) or (ReliesOnSystem some System)
- BusinessDomainItems
- IsDrivenByOrg some Organization

General class axioms

SubClass Of (Anonymous Ancestor)

Instances

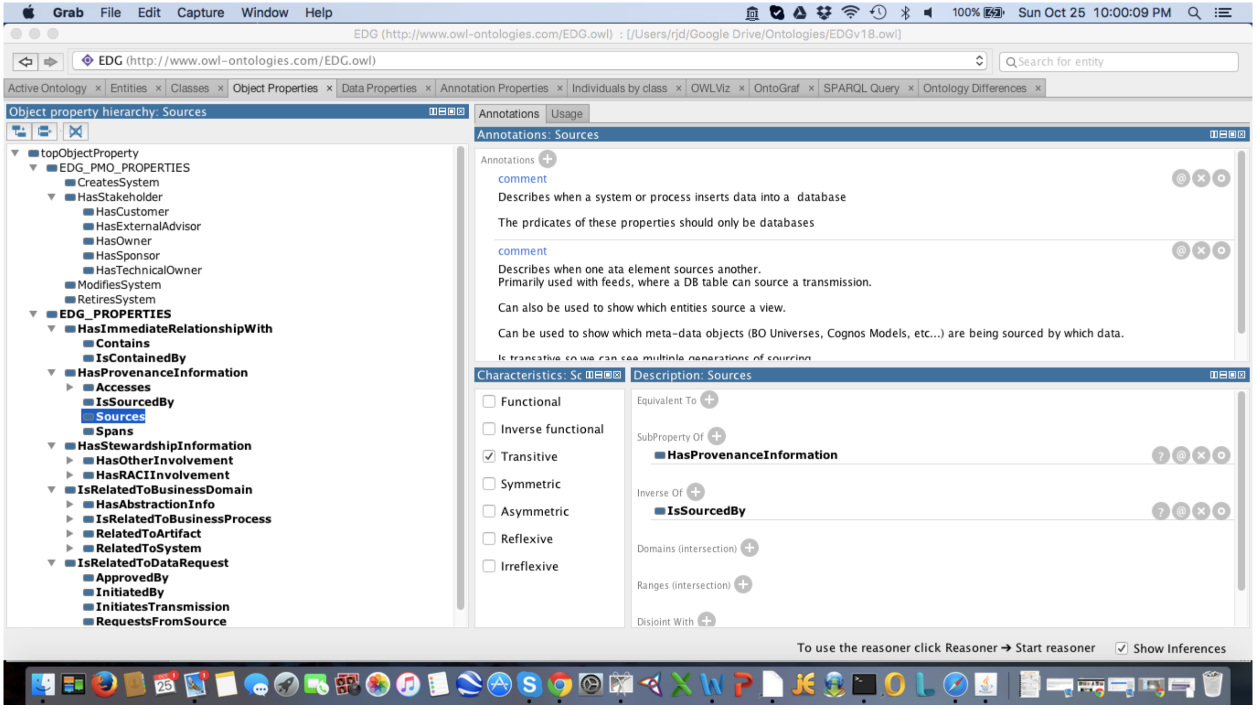
- Process_Budgeting
- Process_CalculateCommissions
- Process_ComissionsAccounting
- Process_EmployeePromotion

Figure 19 Protégé Ontology Class View

In the above figure, the ontology classes are seen on the left side of the screen. They are presented in a "tree view". It is important to note however, that unlike XML, or other tree views, which depict containment, the nature of the classes in an ontology is one based on inheritance. Also important to note is that the class "EDG_PMO_DOMAIN" is describes projects management concepts to integrate with the data management and governance concepts. The reason why it is not in bold is that it is an ontology that is imported into the ontology being viewed in Protégé. As we saw in chapter 3, the ability to import allows for multiple groups to participate in publishing their information, as long as their ontologies are in an accessible location.

On the right side, aside from the section on top with comments and annotations about the highlighted class, the primary function is to highlight restriction on the classes, such as sub-classing, disjoint, equivalence, etc. In addition, the section also shows the instances of the class, otherwise known as “individuals”, as depicted by the purple diamond. While this is quite helpful to be able to view the actual “data” within the ontology, large volumes could be unwieldy to view in this manner.

Similar to the class viewer, is the viewer for object properties. Here, in the same “tree-viewer” format, we can create and manipulated object, as well as data properties (on another tab), as well as set up restrictions. What is different however are the attributes that pertain to properties, such as the check boxes to identify if the properties are functional, transitive, symmetric, etc.



```
topObjectProperty
    EDG_PMO_PROPERTIES
        CreatesSystem
        HasStakeholder
            HasCustomer
            HasExternalAdvisor
            HasOwner
            HasSponsor
            HasTechnicalOwner
        ModifiesSystem
        RetiresSystem
    EDG_PROPERTIES
        HasImmediateRelationshipWith
            Contains
            IsContainedBy
        HasProvenanceInformation
            Accesses
            IsSourcedBy
            Sources
            Spans
        HasStewardshipInformation
            HasOtherInvolvement
            HasRACIInvolvement
        IsRelatedToBusinessDomain
            HasAbstractionInfo
            IsRelatedToBusinessProcess
            RelatedToArtifact
            RelatedToSystem
        IsRelatedToDataRequest
            ApprovedBy
            InitiatedBy
            InitiatesTransmission
            RequestsFromSource
```

Annotations

comment

Describes when a system or process inserts data into a database

The prdicates of these properties should only be databases

comment

Describes when one ata element sources another.
Primarily used with feeds, where a DB table can source a transmission.

Can also be used to show which entities source a view.

Can be used to show which meta-data objects (BO Universes, Cognos Models, etc...) are being sourced by which data.

Is transative so we can see multiple generations of sourcing

□ Functional
□ Inverse functional
☑ Transitive
□ Symmetric
□ Asymmetric
□ Reflexive
□ Irreflexive

Equivalent To

SubProperty Of
HasProvenanceInformation

Inverse Of
IsSourcedBy

Domains (intersection)

Ranges (intersection)

Disjoint With

To use the reasoner click Reasoner → Start reasoner ☑ Show Inferences

Figure 20 Protege Object Property View

Protégé also includes a highly interactive way to view the ontology as well as all the instance data within it by using it’s OntoGraf feature. Here, one can view the various classes, sub classes and instances much like a graph, however, the multiple object properties, or "edges" in this case can be viewed as well, making a highly powerful way to view the various relationships that may bind various entities, which is critical in order to show how enterprise data is regarded within an organization.

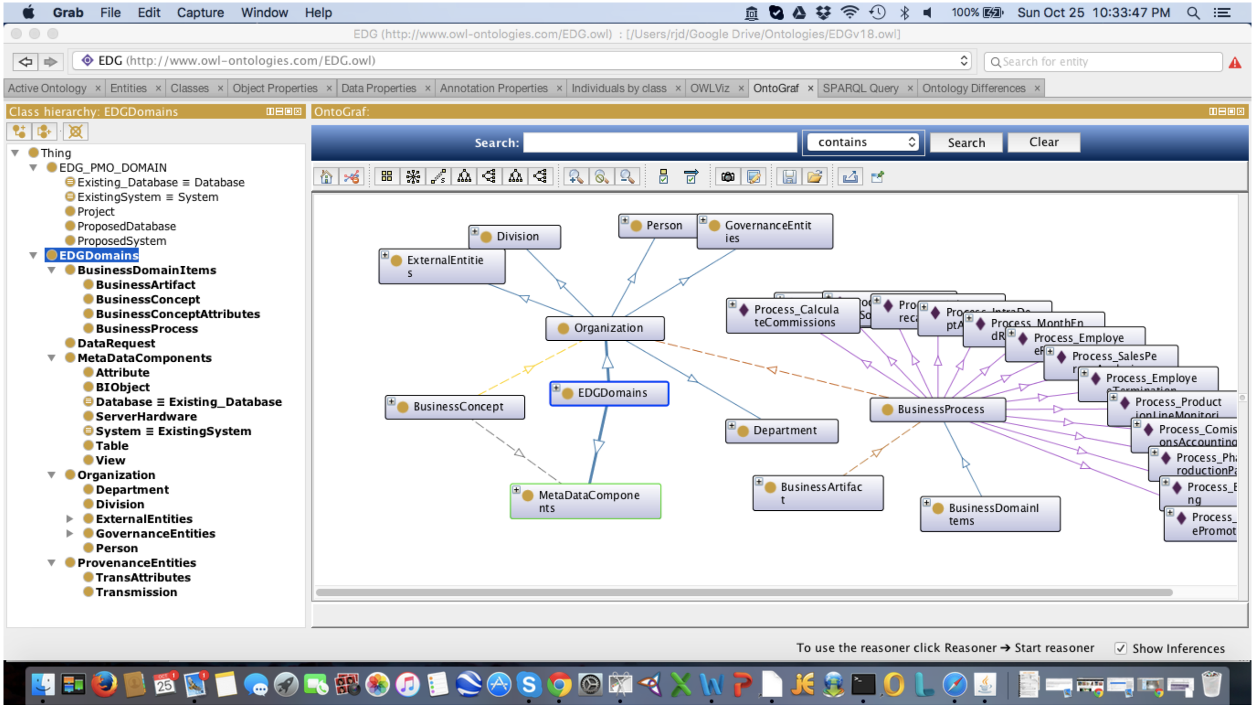


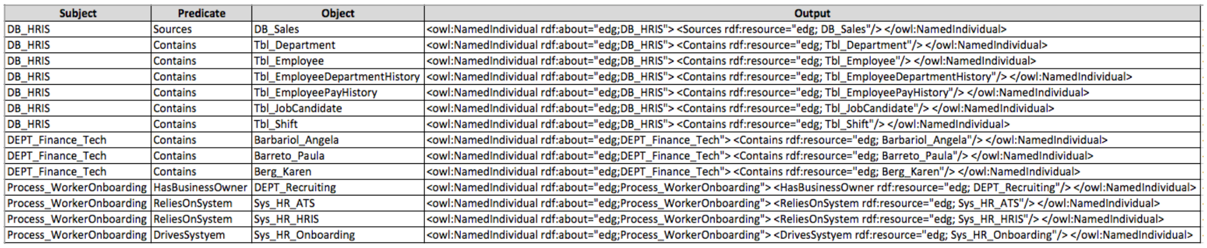
Figure 21 Protégé Ontology Browsing

In addition to Protégé, Gruff [104] is another tool that serves as a graph based triple store browser which is available on a variety of platforms and will be further described in the next chapter.

4.3.2 Development

As with viewers and modelers, there are several different ways to programmatically update data within the ontology. As with our considerations in the prior section, it is important that consistency be maintained when having several different data sets referring to one another or a core enterprise ontology.

As the basis of the data for this research resides in RDF/XML format, it is text base and thereby subject to being manipulated in a variety of ways. In it's most basic form the RDF/XML can be created, transformed, or modified in with standard productivity tools such as MS Excel. While this is primarily a method for smaller sets of stable data, as it is primarily manual, tasks can be automated using Visual Basic for Automation to achieve more scale. The examples below highlight how complex RDF statements can be achieved through Excel formulas, primarily via text handling functions. It is imperative that the relationships and ontology are well defined when transforming this content, as it is manual. In the following example, organizational information is being captured in the form of triples.



Subject	Predicate	Object	Output
DB_HRIS	Sources	DB_Sales	<owl:NamedIndividual rdf:about="edg;DB_HRIS"> <Sources rdf:resource="edg; DB_Sales"/> </owl:NamedIndividual>
DB_HRIS	Contains	Tbl_Department	<owl:NamedIndividual rdf:about="edg;DB_HRIS"> <Contains rdf:resource="edg; Tbl_Department"/> </owl:NamedIndividual>
DB_HRIS	Contains	Tbl_Employee	<owl:NamedIndividual rdf:about="edg;DB_HRIS"> <Contains rdf:resource="edg; Tbl_Employee"/> </owl:NamedIndividual>
DB_HRIS	Contains	Tbl_EmployeeDepartmentHistory	<owl:NamedIndividual rdf:about="edg;DB_HRIS"> <Contains rdf:resource="edg; Tbl_EmployeeDepartmentHistory"/> </owl:NamedIndividual>
DB_HRIS	Contains	Tbl_EmployeePayHistory	<owl:NamedIndividual rdf:about="edg;DB_HRIS"> <Contains rdf:resource="edg; Tbl_EmployeePayHistory"/> </owl:NamedIndividual>
DB_HRIS	Contains	Tbl_JobCandidate	<owl:NamedIndividual rdf:about="edg;DB_HRIS"> <Contains rdf:resource="edg; Tbl_JobCandidate"/> </owl:NamedIndividual>
DB_HRIS	Contains	Tbl_Shift	<owl:NamedIndividual rdf:about="edg;DB_HRIS"> <Contains rdf:resource="edg; Tbl_Shift"/> </owl:NamedIndividual>
DEPT_Finance_Tech	Contains	Barbariol_Angela	<owl:NamedIndividual rdf:about="edg;DEPT_Finance_Tech"> <Contains rdf:resource="edg; Barbariol_Angela"/> </owl:NamedIndividual>
DEPT_Finance_Tech	Contains	Barreto_Paula	<owl:NamedIndividual rdf:about="edg;DEPT_Finance_Tech"> <Contains rdf:resource="edg; Barreto_Paula"/> </owl:NamedIndividual>
DEPT_Finance_Tech	Contains	Berg_Karen	<owl:NamedIndividual rdf:about="edg;DEPT_Finance_Tech"> <Contains rdf:resource="edg; Berg_Karen"/> </owl:NamedIndividual>
Process_WorkerOnboarding	HasBusinessOwner	DEPT_Recruiting	<owl:NamedIndividual rdf:about="edg;Process_WorkerOnboarding"> <HasBusinessOwner rdf:resource="edg; DEPT_Recruiting"/> </owl:NamedIndividual>
Process_WorkerOnboarding	ReliesOnSystem	Sys_HR_ATS	<owl:NamedIndividual rdf:about="edg;Process_WorkerOnboarding"> <ReliesOnSystem rdf:resource="edg; Sys_HR_ATS"/> </owl:NamedIndividual>
Process_WorkerOnboarding	ReliesOnSystem	Sys_HR_HRIS	<owl:NamedIndividual rdf:about="edg;Process_WorkerOnboarding"> <ReliesOnSystem rdf:resource="edg; Sys_HR_HRIS"/> </owl:NamedIndividual>
Process_WorkerOnboarding	DrivesSystyem	Sys_HR_Onboarding	<owl:NamedIndividual rdf:about="edg;Process_WorkerOnboarding"> <DrivesSystyem rdf:resource="edg; Sys_HR_Onboarding"/> </owl:NamedIndividual>

Figure 22 Excel as a Means to Load the Graph

In the above case, the triples are created individually, which is suitable for loading into a triple store. When viewing them however OWL, which is built on XML will nest accordingly, making the presentation more compact, as follows:

```
<owl:NamedIndividual rdf:about="&edg;Process_WorkerOnboarding">
    <rdf:type rdf:resource="&edg;BusinessProcess"/>
    <HasBusinessOwner rdf:resource="&edg;DEPT_Recruiting"/>
    <ReliesOnSystem rdf:resource="&edg;Sys_HR_ATS"/>
```

```
        <ReliesOnSystem rdf:resource="&edg;Sys_HR_HRIS"/>
        <DrivesSystyem rdf:resource="&edg;Sys_HR_Onboarding"/>
    </owl:NamedIndividual>
```

There are numerous tools that exist that convert data into RDF/XML. These tools may be used as part of a one time conversion, or as part of a repeatable process [105].

In addition to tools, Apache Jena [106] is a framework that is based on Java, and provides API's for manipulating both RDF and OWL, as well as a Triple Store platform. It also includes Fuseki, a tool that exposes triple data over HTTP, making it more accessible via web services. We will discuss more of Fuseki in the next section. Finally, Jena also includes a framework for inference, which highlights much of the power of OWL and the semantic web.

As we see in the following code example, the Jena API allows for full automation of all activities around creating, modifying, and querying triple stores. In this example, two new entities, or "resources" are being added to a triple store and the property equivalent class is being applied to link the two in the final "schema.add" statement

```
// State that :system is equivalentClass of :application
Resource resource = schema.createResource(defaultNameSpace + "system");
Property prop = schema.createProperty("http://www.owl-
ontologies.com/EDG.owl#equivalentClass");
Resource obj = schema.createResource("http://www.owl-ontologies.com/EDG-
PMO/application");
schema.add(resource,prop,obj);
```

The following example of a method highlights invoking a reasoner, in this case Pellet on the ontology:

```
private void runPellet( ){
      Reasoner reasoner = PelletReasonerFactory.theInstance().create();
      reasoner = reasoner.bindSchema(schema);
      inferredOnt = ModelFactory.createInfModel(reasoner, EDG_Ontology);
      ValidityReport report = inferredOnt.validate();
}
```

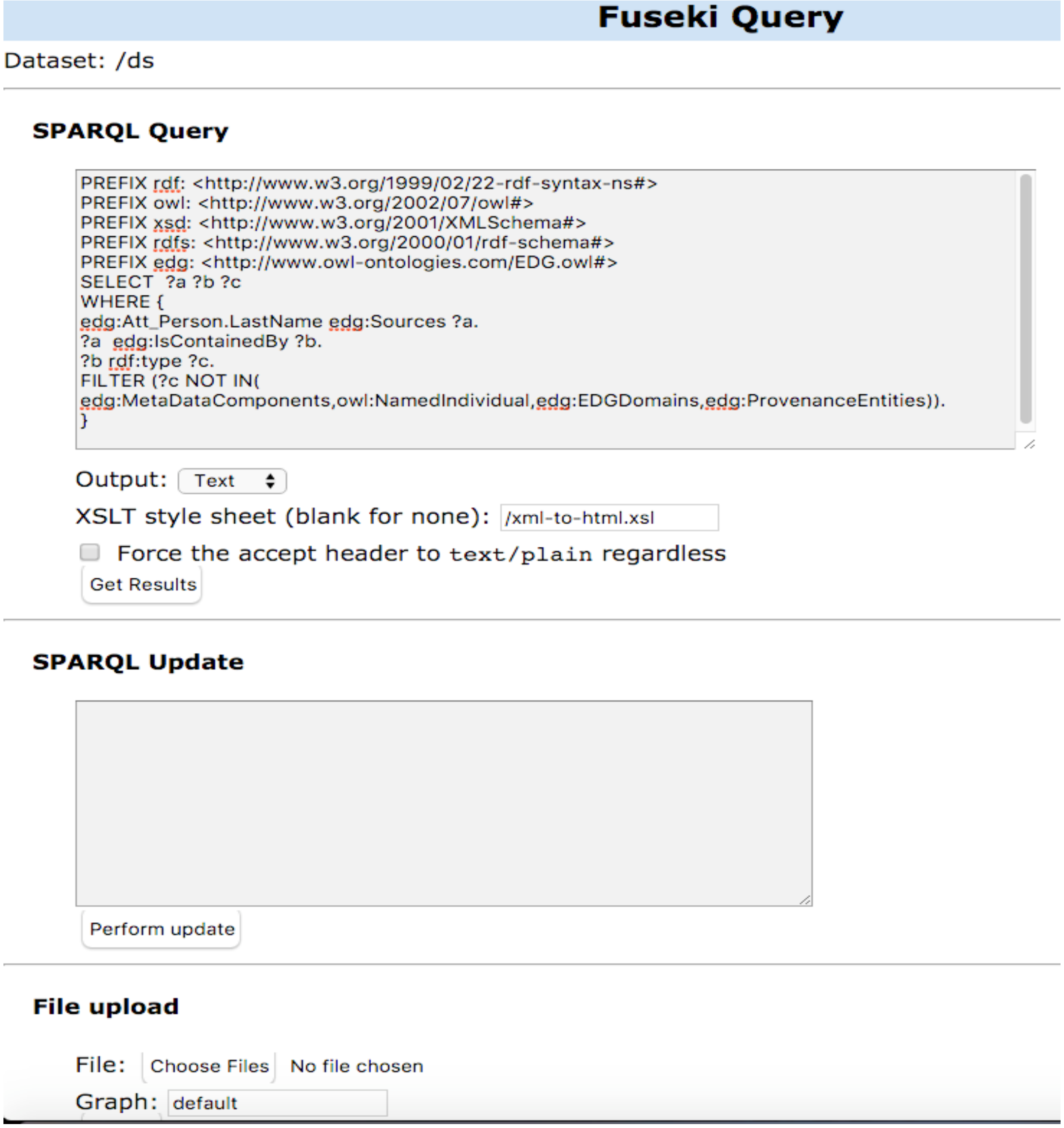
The last example highlights how SPARQL queries can be invoked by the Jena API. This can be quite useful as SPARQL can be used for both reading and updating, and much like SQL, depending on the use case, can put the burden of processing updates upon the triple store, rather than the application.

```
public void myGmailFriends(Model model){
      runQuery(" select ?a ?b ?c WHERE {edg:Att_Person.LastName
      edg:Sources ?a}", model);
}
```

4.3.3 Querying – Fuseki , SPARQL

One of the most significant benefits of this approach is the ability to query the ontology for information regarding the data, projects, and the organization. As we saw in the prior section, there are many ways to visualize the information, however for large graphs with many individuals within the ontology, visualization may be cumbersome to view, or taxing on the system to produce. The primary means of querying ontologies is via SPARQL (SPARQL Protocol and RDF Query Language). To query the system, Apache Fuseki [106] has been utilized to serve as a SPARQL server. Fuseki can run as a stand-alone server, but also a service, or a Java web application. It provides a user interface for server administration and query interaction.

In the case of this research, Fuseki was used as a primary server. Ontology files were saved in OWL/XML format and imported into Fuseki. Files can be imported without imparting any inference and relying on the inference capabilities within Fuseki. While it is feasible to infer a given ontology file and then persist the fully inferred version, it's size can increase drastically in the process, as all axioms will be explicitly asserted in the new file. The screenshot below shows the main query screen within Fuseki:



Fuseki Query

Dataset: /ds

SPARQL Query

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX edg: <http://www.owl-ontologies.com/EDG.owl#>
SELECT  ?a ?b ?c
WHERE {
edg:Att_Person.LastName edg:Sources ?a.
?a  edg:IsContainedBy ?b.
?b rdf:type ?c.
FILTER (?c NOT IN(
edg:MetaDataComponents,owl:NamedIndividual,edg:EDGDomains,edg:ProvenanceEntities)).
}
```

Output: Text

XSLT style sheet (blank for none): /xml-to-html.xsl

□ Force the accept header to text/plain regardless

SPARQL Update

File upload

File: Choose Files No file chosen

Graph: default

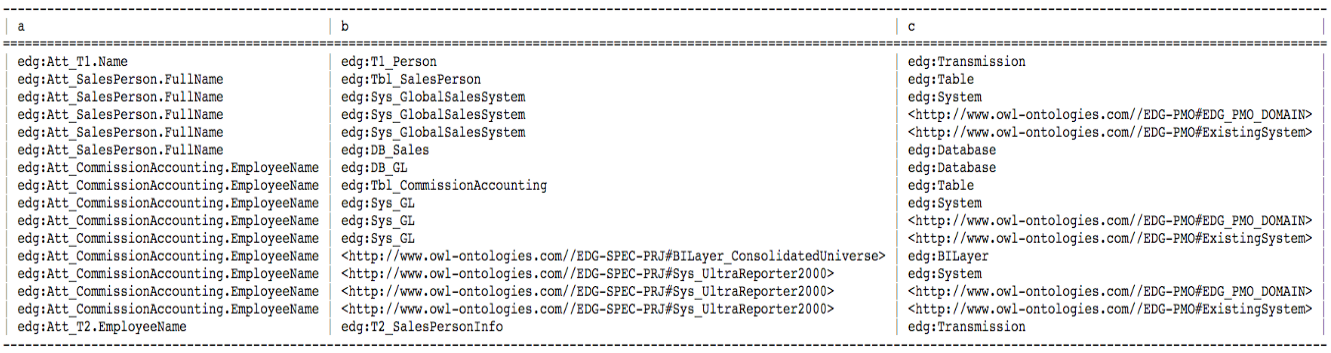
Figure 23 Fuseki example

Here, we see that the contents of a SPARQL query can be entered into the control panel. Updates can also be entered from this screen. The user also has the option to specify how the query results should be retrieved, either in text, JSON, XML, or in CSV format. In the above query example the following result excerpts are in JSON and text formats respectively:

JSON:

```
{   "head": {"vars": [ "a" , "b" , "c" ]   } ,   "results": {"bindings": [
{"a": { "type": "uri" ,
"value": "http://www.owl- ontologies.com/EDG.owl#Att_T1.Name" } ,
"b": { "type": "uri" ,
"value": "http://www.owl-ontologies.com/EDG.owl#T1_Person" } ,
"c": { "type": "uri" ,
"value": "http://www.owl-ontologies.com/EDG.owl#Transmission" }
}
```

TEXT:



```
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
| a                                         | b                                                                          | c                                                       |
====================================================================================================================================================================================
| edg:Att_T1.Name                           | edg:T1_Person                                                              | edg:Transmission                                        |
| edg:Att_SalesPerson.FullName              | edg:Tbl_SalesPerson                                                        | edg:Table                                               |
| edg:Att_SalesPerson.FullName              | edg:Sys_GlobalSalesSystem                                                  | edg:System                                              |
| edg:Att_SalesPerson.FullName              | edg:Sys_GlobalSalesSystem                                                  | <http://www.owl-ontologies.com//EDG-PMO#EDG_PMO_DOMAIN> |
| edg:Att_SalesPerson.FullName              | edg:Sys_GlobalSalesSystem                                                  | <http://www.owl-ontologies.com//EDG-PMO#ExistingSystem> |
| edg:Att_SalesPerson.FullName              | edg:DB_Sales                                                               | edg:Database                                            |
| edg:Att_CommissionAccounting.EmployeeName | edg:DB_GL                                                                  | edg:Database                                            |
| edg:Att_CommissionAccounting.EmployeeName | edg:Tbl_CommissionAccounting                                               | edg:Table                                               |
| edg:Att_CommissionAccounting.EmployeeName | edg:Sys_GL                                                                 | edg:System                                              |
| edg:Att_CommissionAccounting.EmployeeName | edg:Sys_GL                                                                 | <http://www.owl-ontologies.com//EDG-PMO#EDG_PMO_DOMAIN> |
| edg:Att_CommissionAccounting.EmployeeName | edg:Sys_GL                                                                 | <http://www.owl-ontologies.com//EDG-PMO#ExistingSystem> |
| edg:Att_CommissionAccounting.EmployeeName | <http://www.owl-ontologies.com//EDG-SPEC-PRJ#BILayer_ConsolidatedUniverse> | edg:BILayer                                             |
| edg:Att_CommissionAccounting.EmployeeName | <http://www.owl-ontologies.com//EDG-SPEC-PRJ#Sys_UltraReporter2000>        | edg:System                                              |
| edg:Att_CommissionAccounting.EmployeeName | <http://www.owl-ontologies.com//EDG-SPEC-PRJ#Sys_UltraReporter2000>        | <http://www.owl-ontologies.com//EDG-PMO#EDG_PMO_DOMAIN> |
| edg:Att_CommissionAccounting.EmployeeName | <http://www.owl-ontologies.com//EDG-SPEC-PRJ#Sys_UltraReporter2000>        | <http://www.owl-ontologies.com//EDG-PMO#ExistingSystem> |
| edg:Att_T2.EmployeeName                   | edg:T2_SalesPersonInfo                                                     | edg:Transmission                                        |
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
```

Figure 24 Fuseki Query Results

While there are query capabilities presented in Protégé, the tool is primarily a desktop application used for viewing and editing ontologies. While it can function for troubleshooting and individual development, a dedicated SPARQL server provides a more scalable solution. To that end, TDB [106], also part of the Apache Jena project is a high performance triple store that tailored for high volume queries. Used in conjunction with Apache Inference API, could provide a highly scalable and economical platform for a large-scale organization.

In addition, the tool Gruff, a graph database visualizer, which sits atop the Allegro Graph platform, are both products from Franz, Inc. [104] which are highly scalable. Gruff allows the user to graphically create queries, much like a traditional business intelligence tool, in which the SPARQL syntax is generated. Additionally, it provides a great deal of flexibility when displaying and manipulating large graphs. The Allegro Graph platform is also highly scalable and can be hosted in the cloud for additional scalability across large organizational footprints. Examples of Gruff can be seen in the following chapter in which query results are analyzed for various use cases.

4.3.4 Frameworks - Callimachas

Finally, while we have discussed various components to constructing a solution that leverages Linked Data and Ontology, there are frameworks that allow for the construction of large-scale applications based on these principles. One such framework is Callimachus [107]. Callimachus is a type of application server that allows for software to be built to take advantage of linked open data. It is open source and provides browser-based tools for developers to build linked data applications. Callimachus has several major features [75]:

- A template system to automatically generate web pages for each member of an OWL class. OWL classes are technically either equivalent to or subclasses of RDF Schema classes (depending on the OWL profile used), but for our purposes you can think of them as being equivalent.

- An ability to retrieve data at runtime and convert it to RDF.
- An ability to associate SPARQL queries with URLs, to parameterize those queries, and to use their results with charting libraries.
- An implementation of persistent URLs (PURLs).
- A structured writing system based on DocBook and including a visual editing environment.

The benefit here is that an enterprise application or web site can be created based on this framework that sourced information from many different areas and in different formats. While it's primary goal was to bring together information as diverse as what's on the internet, having the additional flexibility while be very beneficial in allowing multiple groups across large organizations store their information in ways that suits them best. This will help foster maximum adoption across the enterprise of the enterprise data ontology, thereby increasing it's probability of success.

Chapter 5

Solution Setup and Testing

5.1 Testing setup and Environment

In this chapter, we will look at the set up of the environment used to simulate and execute the use cases used to validate the efficacy of this research. Implementation methods will be reviewed, in addition to the test suite of data and how it was populated. We will then review several key enterprise use cases in which visibility into the data environment would be imperative.

5.1.1 Tools and Environment

As mentioned in the prior chapter, the nature of linked open data, ontology, and a wide ecosystem of tools, allows for a very flexible and economical approach to cataloging enterprise data assets, organizational entities, and relationships across them. This research was no different, in that, several different tools and techniques were employed in its construction. The following tools were employed:

- Protégé 5.1: Primary tool for designing and constructing the ontologies and initial seeding of individual data.
- Gruff 6.0.1 & Allegrograph 5.1 Server: Triple store and graph viewing and querying tool for navigating and reporting on the data, as well as manipulating elements of the ontology.
- Apache Fuseki 0.2.7: Primary tool for executing SPARQL queries for both reporting as well as updating the triple stores.

- Microsoft Excel: Used for manipulating relational data in order to get into a OWL format.
- Apache Jena: Used in conjunction with Java for automating some of the loading of the data into a triple store.

The tools were all run on an OS X platform, with the ontologies being hosted on a Linux webserver to simulate the ability to publish changes. Ontologies were stored in OWL/XML format. For ontology visualization and manipulation, Protégé was used for it's ease of use in editing the classes, properties, and individuals, as well as viewing all the axioms. For complex queries and visualization, the ontologies were imported in to a triple store, Alegrograph and queried/viewed with Gruff. Complex SPARQL queries were also handled with Jena Fuseki. The tools performed well (less than 1 second response time) under the current circumstances, however when visualizing the graph structures, as the database grew to approximately 20,000 triples, the current hardware was strained and times to render the graphs were between 10 and 20 seconds.

5.1.2 Data Setup

Test data was created to provide a representation of a diverse organization, reflecting the key divisions and their subsequent departments to run it. Additionally, a diverse set of systems, employees, and other data provenance and governance information was created in order to show the dynamic interactions between data and people.

Three ontologies were created in order to illustrate the distributed nature of ontology and linked data. The first is the primary enterprise data graph. This ontology contains all of the primary organizational and structural master data required to serve as a foundation across the organization. The next ontology highlights Program Management Office (PMO) information in order to illustrate how various projects can span data and organization al entities. The third and a departmentally specific ontology that contains information regarding the systems landscape of that group only. In the case of this testing, the ontology was based on the specific systems that are focused for a Compliance department. This use case is to highlight the ability to publish departmental or local level information that will be integrated into a larger picture, thereby empowering groups at all levels of the organization.

It is also important to note that while the ontologies used in this testing vary as far as individuals go, they share the same classes and properties. While subsidiary ontologies are free to extend on the set of classes and properties, and it is true that by creating equivalent classes to maintain semantic consistency, in practice, in the closed environment of an organization, it reduces maintenance and upfront development efforts to maintain ontological consistency by establishing a core vocabulary that is to be adopted by all.

For this research, the following metrics pertain to the test data that was created:

Total Axioms:	3230
Total Classes:	35
Total Subclasses:	74
Total Individuals:	901
Total Object Properties:	67
Total Axioms (Inferred):	18,349

In terms of organizational structure, the ontology represented a sampling of an organization with the following metrics:

People:	232
Departments:	26
Divisions:	8
Systems:	20
Business Processes:	16

Given the size of the sample data, it is reasonable that the inferred number of triples will range in the millions, given an organization with tens of thousands of employees and thousands of systems, databases, and integrations. While the incremental approach to this framework is will allow for maximum participation, additional resources can be focused on performance and tuning.

5.2 Use Cases

The following use cases highlight ways in which this approach can be utilized in a corporate environment. Having the information regarding physical aspects of data tends to only tell a portion of the story needed to make informed decisions on governing that data. Coupling it with organizational and provenance information adds a richness that can lead to better data governance.

5.2.1 Data Lineage & Impact Analysis of Field Change

One of the most ubiquitous use cases for managing information and having a strong data awareness is knowing what the impact is when a piece of data changes. Often times, a very universal data element, such as Employee ID is used across organizations to see many systems. Making a change to the field, such as data type or field length changes can have vast impacts across the organization.

In this example, we want to analyze the touch points and dependencies on fields that pertain to “Last Name” within the organization. In this case, the following SPARQL query can yield that information:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX edg: <http://www.owl-ontologies.com/EDG.owl#>

SELECT  ?impactedObject ?parentObject  ?steward
WHERE {
edg:Att_Person.LastName edg:Sources+ ?impactedObject.
?impactedObject  edg:IsContainedBy ?parentObject.
?parentObject edg:HasStewardshipInformation ?steward.
}
```

In the above query, we make use of some key aspects of the flexible nature of ontology and a graph structure. First, the expressive nature of otology, and OWL in particular allows us to specify data in as much or as little detail as we can and inference can assist in bridging the gaps. In this case, the object property *IsContainedBy* is defined as an inverse property to *Contains*. This allows the users in a corporate environment to only need to specify one direction of the relationship. Additionally, the transitive nature of the Resource Definition Framework allows us to have the object property *Sources* defined as such, so its relationship can span generations. This is denoted by the “+” sign after the property. This is particularly important when highlighting concepts of lineage.

In the screenshot below, from Gruff, it can be seen that the last name field sources four other objects directly (denoted by the green arrows). In addition, overall system impacts and stakeholders can be viewed at the same time. Here, we see that the Last Name field sources fields in other systems, which in this case are denoted as Systems (in light blue) or other transmissions (in purple). The departmental stakeholders of those downstream dependencies are also denoted in yellow.

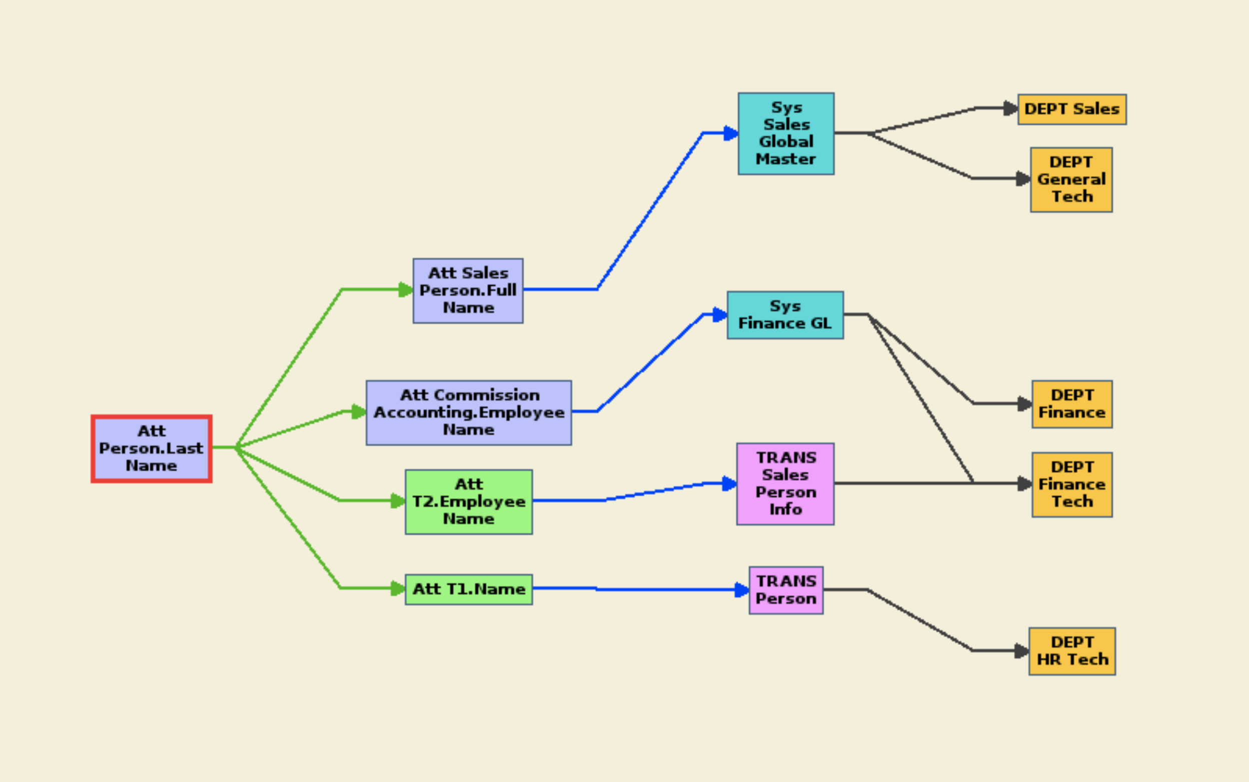


Figure 25 Impact Analysis Example

5.2.2 Critical User Analysis

Being able to identify the critical users of a system is essential. This is important not only for planning for change, but also should there be system issues that require an emergency notification. For large systems with many disparate users, the user base can be spread out across the organization. Often times, teams must refer to the system audits to find who is frequently accessing the systems to identify the critical stakeholders. The following example extends on the prior example in that it illustrates all the critical users of the systems that are impacted by a change to the Last Name field.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX edg: <http://www.owl-ontologies.com/EDG.owl#>
SELECT  ?f
WHERE {
edg:Att_Person.LastName edg:Sources ?a.
?a  edg:IsContainedBy ?b.
?b edg:HasStewardshipInformation ?d.
#find the relationships that exist between those with stewardship info and
the affected system components
#Only return critical users and the tech owners
?d ?e ?b.
FILTER (?e IN(edg:IsCriticalUserOf)).
?d edg:Contains ?f.
?f rdf:type ?g.
FILTER (?g = edg:Person).
}
```

5.2.3 Business Process Analysis

The next use case illustrates the situation where it is necessary to know what the span of a given process is within an organization. This highlights the fact that in order to properly govern data it is essential to understand the how it relates to the organizational structure as well as to its operations. Often this linkage is not formally recorded anywhere therefore, having the ability to empower the individuals who focus on the day to day operations is imperative to gathering rich information. Here, we see what the span of the Worker Onboarding Process is within an organization.

In large organizations, the process of onboarding a new employee touches across several systems, such as the Applicant Tracking System, Recruiting websites and portals, the core HRIS system, the compensation system, security and provisioning systems, etc. When a company is faced with corporate actions such a merger, or expansion internationally, these processes and systems may be subject to change. The following visualization and subsequent SPARQL query illustrate how a flexible framework allows for linking of as much process information as the users enter with the current data landscape.

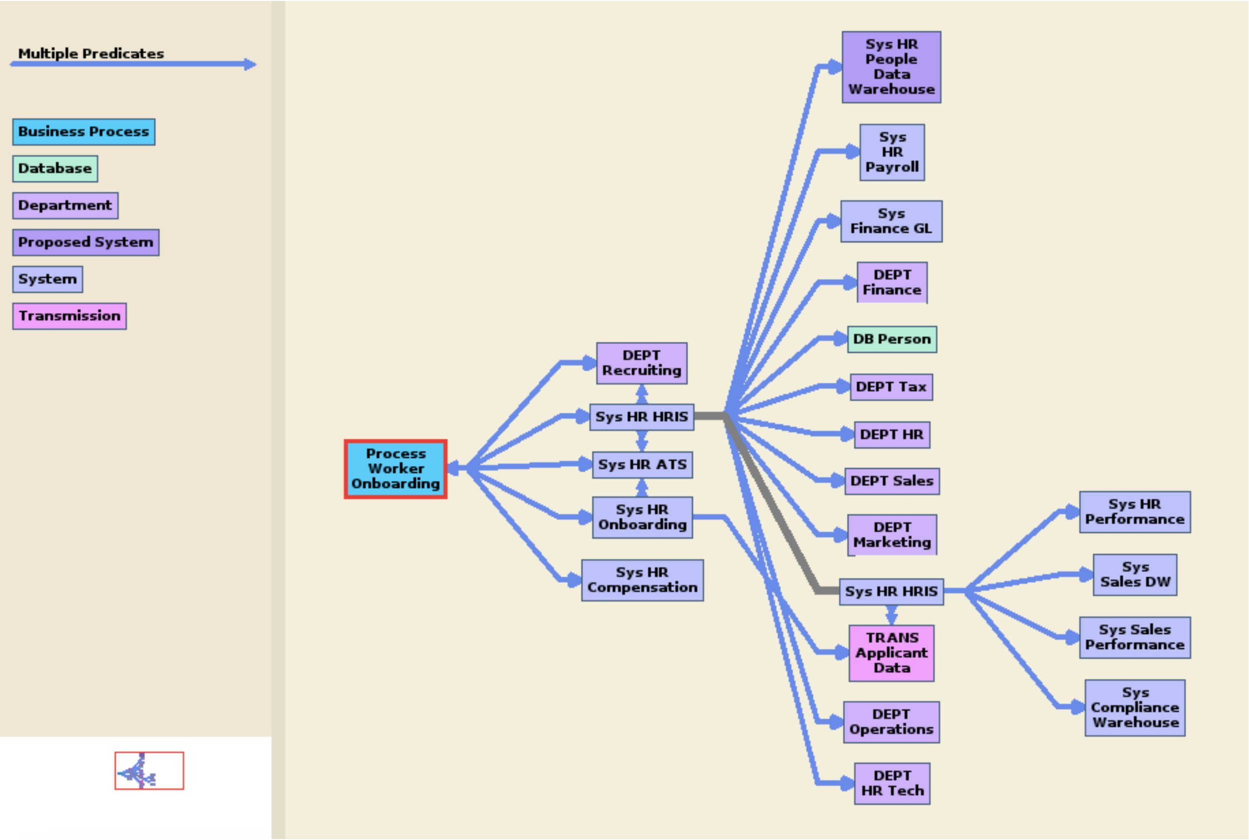


Figure 26 Business Process Example

```
select ?node_variable_1 ?predicate_variable_2 where

{<http://www.owlontologies.com/EDG.owl#Process_WorkerOnboarding>
?predicate_variable_2 ?node_variable_1.}
```

The above SPARQL statement, while not complex, was generated from the Gruff and can readily illustrate high-level touch points for core organizational processes. Visualization tools, much like their Business Intelligence counterparts allow for real time drilling and manipulation of the findings.

5.2.4 Project Analysis

The next example shows how this framework can be expanded to include linking the organizational data environment to projects within the organization. This is important in identifying the impact of such projects so there can be appropriate planning for integration testing, communications planning, as well as putting together a technology adoption strategy for any new development. Additionally, some of the individuals in this query reside in another ontology, as seen by the additional namespace "pmo:". This highlights how the framework can be distributed in such a manner that centralized groups within and organization, such as the Program Management Office, which is usually charged with maintaining the current portfolio of work for a given area can publish their own information about what given data assets pertain to particular processes.

The screenshot below, and subsequent SPARQL illustrate how the "People Reporting" project which manifests itself in the build out of a "Proposed System" called the "People Data Warehouse" touches existing systems and organizational entities. This can be very helpful in identifying all the appropriate stakeholders for a given project as early as possible, thereby reducing surprise and subsequent last minute work.

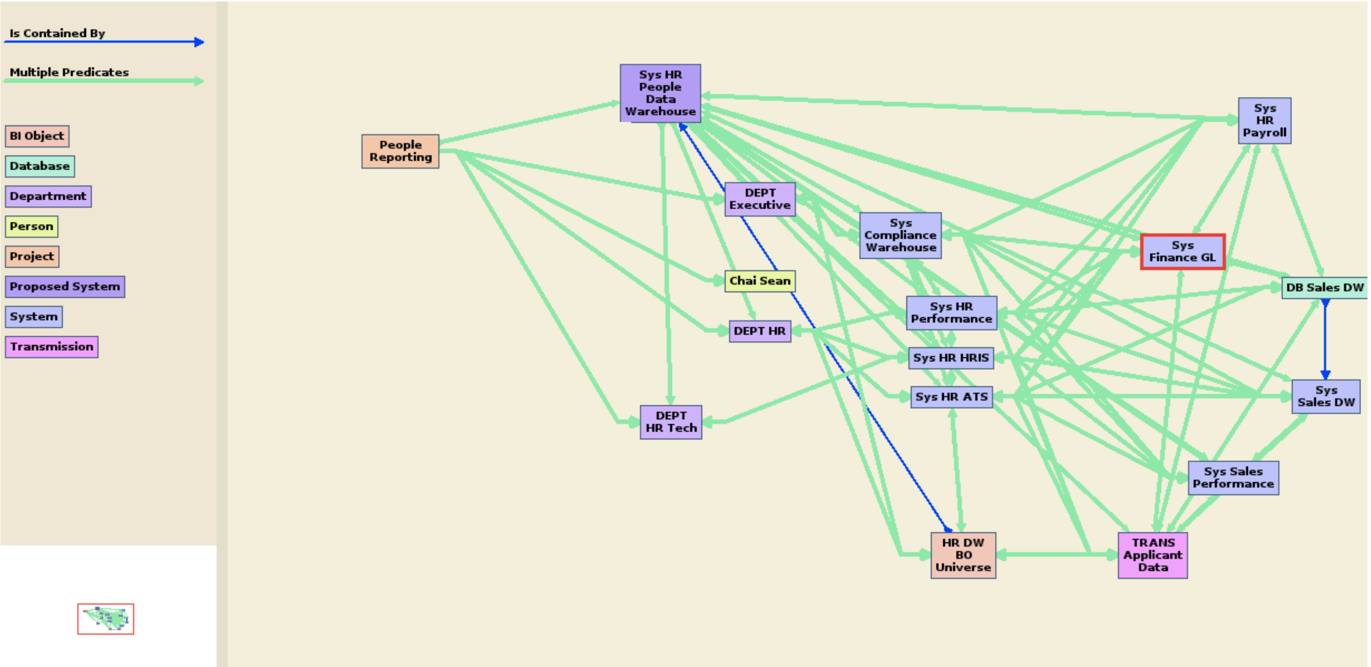


Figure 27 Project Analysis Example

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX owl: <http://www.w3.org/2002/07/owl#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

PREFIX edg: <http://www.owl-ontologies.com/EDG.owl#>

PREFIX pmo: <http://www.owl-ontologies.com//EDG-PMO#>


SELECT  ?a ?d ?c

WHERE {

pmo:ProjectFinCRM ?b ?c.

#FILTER (?b = pmo:DemiseSystem).

#?a edg:HasStewardshipInformation* ?c.

?a ?d ?c
}
```

5.3 Conclusion

In this chapter, we have seen the fundamental implementations of this research across several different toolsets and with varying data that is representative of organizational structure, initiatives, data assets, etc. As mentioned, the flexible nature of the framework allows for companies to adopt as much as they are able to at any point in time. The primary architectural decisions to make will be driven by the scale at which the organization adopts this approach, and the raw size of the company as well. In the cases of large volumes of data and at low levels of detail, more robust triple stores, such as Allegrograph, or Neo4j may be needed in order to provide responsive browsing and manipulating of the data. In those cases, still groups can embrace a “pay as you go” [108] approach to adoption.

Chapter 6

Solution Summary and Future Work

6.1 Summary

We have seen over the past decade, that data has been an increasingly valuable asset with organizations[1]. Because of this, it is imperative that companies adopt policies and tools to govern this asset as they would any other. Data can be thought of as the “oil that flows through the machinery” of the organization. It is this distributed and dynamic nature of data however, unlink other corporate assets that require a heightened level of awareness as to how it relates to the organization and other technology assets in order to effectively govern it. This need for governance, and subsequent awareness is further increased as we see the complex nature of business today. With mergers and acquisitions, globalization, distributed work, outsourcing, corporate structures are getting more distributed and more complex. Additionally, regulations such as BASEL II, Sarbanes Oxley, and Dodd Frank have increased the need for companies to have a keen awareness as to their data, and the processes that can create, modify, and destroy it. Adding to the complexity is the fact that the decreased cost for storage, and the ability to instrument and audit more activities, and the data generating scale of mobile applications, have led to an explosion in the amount of data being captured and our ability to store it. Once stored, companies are looking for ways to gain advantages and insights in tapping those data assets in ways to improve operations, as well as flesh out opportunities.

As stated in chapter two, the flow of data throughout an organization has been likened to a supply chain, in that raw ingredients become processed, combined, and manipulated along the way to becoming a finished product. In the case of data, it can move throughout it's lifecycle, first being created, then being updated, copied, moved, or transported throughout the organization to eventually be archived or destroyed. This transient nature, with many touch points, parties, and linkages lends itself to a graph structure to model it's lifespan. Additionally, because of the rich relationships that need to be captured in order to reflect how data relates differently to different parties, systems, and initiatives, ontology is a suitable approach to model these complex relationships. As there needs to be significant participation from the organization for the richness of the relationships to emerge, concepts of linked data [29] as proposed by Tim Berns-Lee allow for a lower barrier to entry in having significant adoption across an enterprise setting. When all is taken in conjunction with the fact that there are many rich open source tools and methods for storing, collecting, and visualizing this disparate data, we have a solution that is not only scalable and flexible, but also economical. As data centric enterprise projects have been fraught with difficulties in implementation and adoption, It is these lower barriers to entry that make this research a viable framework for organizations seeking to increase their data awareness and subsequent data governance efforts.

6.2 Future Research

Much of the potential future research based on this thesis can be focused in the areas of industrializing the framework. As we have seen in chapter 4 where various methods of implementation were discussed, capturing data such as the organizational structure and the physical data base elements prove to be straight forward as they can usually be gotten from a system of record for the companies structure (generally the HR or Financial Systems), as well as the database catalogs for all of the respective systems. Capturing elements, such as unstructured data will require additional intelligence to parse out that information and tag its elements correctly. Document contents and email data are the largest components to this information. Also, data transmission, such as flat file feeds from one system to another, web service calls, and data exchange are an area where research can extend the scalability of this framework so less manual interaction is required to populate said interchanges and link them to their respective sources and targets. An example of such an instance would be the case where a flat file of information, for example all current employees and some additional attribute information are being sent from the core HR system to another system. In some cases, an SQL query is the source of the file that will be transported. Having logic that will read the query and derive the content from both the SELECT and WHERE clauses, in addition handling joins and nested sub queries, while mapping that to the existing ontology contents would increase the scalability of such a solution.

Conversely, as such an ontology can be built out to contain elements of abstraction, much like a business intelligence system's meta data layers, query generation from the ontology components can be another interesting direction that can lead to increased usefulness in an organization.

Increasing adoption across the enterprise is an additional area for future research. While the decentralized nature of linked open data provide for distributed publishing, a robust enterprise application or set of processes with a common look and feel can increase the consistency of the data across the enterprise. In these cases, providing a user experience that is ubiquitous and accessible to the users throughout the day, such as embedding such functionality on the desktop, or in the corporate intranet, will allow everyone to be able to update (if they have the appropriate access) and view any of the data landscape across the organization. Expansion in this area will further increase adoption of the tool and increased richness.

Further research can focus on further integration with corporate systems that can be indicative of the relationships across people within the organization. For example, email traffic can be retrieved and placed into a directed graph structure. Internal and corporate social networking platforms, such as Yammer can also be mined to provide network information about how people relate to one another. This can in turn be coupled with organizational information and overlaid with governance information as to what people's roles are throughout the organization and how it overlays with the data they require to do their job.

Appendix

A. SPARQL QUERIES:

Example 1

```
# The following query uses transitive properties in order to identify all the affected
components of the

# LastName field in order to perform an impact analysis.

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX owl: <http://www.w3.org/2002/07/owl#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

PREFIX edg: <http://www.owl-ontologies.com/EDG.owl#>

SELECT  ?a ?b ?c

WHERE {

edg:Att_Person.LastName edg:Sources ?a.

?a  edg:IsContainedBy ?b.

?b rdf:type ?c.

FILTER (?c NOT IN(
edg:MetaDataComponents,owl:NamedIndividual,edg:EDGDomains,edg:ProvenanceEntities)
).

}
```

Example 2

```
# Identify all of the critical users that are stakeholders of that field

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX owl: <http://www.w3.org/2002/07/owl#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX edg: <http://www.owl-ontologies.com/EDG.owl#>

SELECT  ?impactedObject ?parentObject  ?steward

WHERE {

edg:Att_Person.LastName edg:Sources ?impactedObject.

?impactedObject  edg:IsContainedBy ?parentObject.

?parentObject edg:HasStewardshipInformation ?steward.

}

#find the relationships that exist between those with stewardship info and the affected system components

#Only return critical users and the tech owners

?d ?e ?b.

FILTER (?e IN(edg:IsTechnicalOwnerOf, edg:IsCriticalUserOf))

}
```

Example 3

```
# Identify all of the critical users that are stakeholders of that field

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX owl: <http://www.w3.org/2002/07/owl#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

PREFIX edg: <http://www.owl-ontologies.com/EDG.owl#>

SELECT  ?f

#SELECT  ?a ?b  ?d ?e ?f

WHERE {

edg:Att_Person.LastName edg:Sources ?a.

?a  edg:IsContainedBy ?b.
?b edg:HasStewardshipInformation ?d.

#find the relationships that exist between those with stewardship info and the affected system components

#Only return critical users and the tech owners

?d ?e ?b.

FILTER (?e IN(edg:IsCriticalUserOf)).

?d edg:Contains ?f.

?f rdf:type ?g.

FILTER (?g = edg:Person).

}
```

#Example 4

```
# Identify all of the stakeholders of a system that will be demised from a particular project

#This query brings together data from multiple ontologies

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX owl: <http://www.w3.org/2002/07/owl#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

PREFIX edg: <http://www.owl-ontologies.com/EDG.owl#>

PREFIX pmo: <http://www.owl-ontologies.com//EDG-PMO#>

SELECT  ?a ?d ?c

WHERE {

pmo:ProjectFinCRM ?b ?c.

FILTER (?b = pmo:DemiseSystem).

?a edg:HasStewardshipInformation* ?c.

?a ?d ?c

}
```

Example 5

```
# The following query uses a data type property to identify what attributes are sensitive data.
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX edg: <http://www.owl-ontologies.com/EDG.owl#>
SELECT  ?a ?b ?c
WHERE {
edg:Tbl_Person edg:Contains ?a.
?a edg:SensetiveData ?b
}
```

References

[1] Z. Panian, “Some practical experiences in data governance,” *World Academy of Science, Engineering, Technology, and Management, vol. 62, no.1*, 2010, pp.939-946.

[2] V. Khatri and C. V. Brown, “Designing data governance,” *Communications of the ACM*, vol. 53, no. 1, pp. 148–152, Jan. 2010.

[3] S. Sarsfield, *The Data Governance Imperative*. IT Governance Ltd, 2009.

[4] D. International, “The DAMA Guide to the Data Management Body of Knowledge,” *The DAMA Guide to the Data Management Body of Knowledge*, Feb. 2010.

[5] D. S. Sayogo, T. A. Pardo, and P. Bloniarz, *Information sharing and financial market regulation: understanding the capability gap*. New York, New York, USA: ACM, 2012, pp. 123–131.

[6] T. A. Pardo, J. R. Gil-Garcia, and G. B. Burke, “Sustainable Cross-Boundary Information Sharing,” in *Digital Government*, vol. 17, no. 21, H. Chen, L. Brandt, V. Gregg, R. Traunmüller, S. Dawes, E. Hovy, A. Macintosh, and C. A. Larson, Eds. Boston, MA: Springer US, 2008, pp. 421–438.

[7] T. A. Pardo, D. S. Sayogo, and D. S. Canestraro, “Computing and information technology challenges for 21st century financial market regulators,” presented at the EGOV'11: Proceedings of the 10th IFIP WG 8.5 international conference on Electronic government, Berlin, Heidelberg, 2011, vol. 6846, no. 17, pp. 198–209.

[8] T. Friedman, M. A. Beyer, and E. Thoo, “Magic quadrant for data integration tools,” *Gartner RAS Core Research*, p. 35, Nov. 2010.

[9] J. Radcliffe, “Magic quadrant for master data management of customer data,” *Gartner Research*, Jan. 2009.

[10] L. Moreau, J. Freire, J. Futrelle, and R. E. McGrath, “The open provenance model: An overview,” *Provenance and Annotation of Data and Processes. Springer Berlin Heidelberg*, pp. 323–326, 2008.

[11] P. Weill and J. W. Ross, *IT Governance: How Top Performers Manage IT Decision Rights for Superior Results*. Boston: Harvard Business School Press, 2004.

[12] P. Aiken, M. Gillenson, X. Zhang, and D. Rafner, “Data Management and Data Administration: Assessing 25 Years of Practice,” *Journal of Database Management*, vol. 22, no. 3, pp. 24–45, Jul. 2011.

[13] T. Evgeniou, “Information integration and information strategies for adaptive enterprises,” *European Management Journal*, vol. 20, no. 5, pp. 486–494, 2002.

[14] H. S. Jagdev and K. D. Thoben, “Anatomy of Enterprise Collaborations,” *Production Planning & Control*, vol. 12, no. 5, pp. 437–451, Nov. 2010.

[15] W. W. Eckerson and R. P. Sherman, “Strategies for managing spreadmarts,” *Business Intelligence Journal*, vol. 13, no. 1, pp. 23–24, 2008.

[16] P. Aiken, M. D. Allen, B. Parker, and A. Mattia, “Measuring Data Management Practice Maturity: A Community's Self-Assessment,” *Computer*, vol. 40, no. 4, pp. 42–50, Apr. 2007.

[17] E. M. Trauth, “The evolution of information resource management,” *Information & Management*, vol. 16, no. 5, pp. 257–268, May 1989.

[18] M. L. Gillenson, “Database Administration at the Crossroads: The Era of End-User-Oriented, Decentralized Data Processing,” *Journal of Database Management (JDM)*, vol. 2, no. 4, pp. 1–11, 1991.

[19] J. A. Vayghan, S. M. Garfinkle, C. Walenta, D. C. Healy, and Z. Valentin, “The internal information transformation of IBM,” *IBM Systems Journal*, vol. 46, no. 4, pp. 669–683, 2007.

[20] G. Singh, S. Bharathi, A. Chervenak, E. Deelman, C. Kesselman, M. Manohar, S. Patil, and L. Pearlman, *A Metadata Catalog Service for Data Intensive Applications*. IEEE, 2003, pp. 33–33.

[21] T. K. Das and M. R. Mishra, “A study on Challenges and Opportunities in Master Data Management,” *International Journal of Database Management Systems IJDMS*, vol. 3, no. 2, 2011.

[22] J. M. Juran, *Leadership for Quality: An Executive Handbook*. The Free, 1989.

[23] L. P. English, *Improving data warehouse and business information quality*. J. Wiley and Sons, 1999.

[24] S. Watts, G. Shankaranarayanan, and A. Even, “Data quality assessment in context: A cognitive perspective,” *Decision Support Systems*, vol. 48, no. 1, pp. 202–211, Dec. 2009.

[25] P. Russom, “TDWI Checklist Report: Cost Justification for Metadata Management,” *The Data Warehouse Institute*, pp. 1–10, Aug. 2010.

[26] T. E. Elliott, J. H. Holmes, A. J. Davidson, P.-A. La Chance, A. F. Nelson, and J. F. Steiner, “Data Warehouse Governance Programs in Healthcare Settings: A Literature Review and a Call to Action,” *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, vol. 1, no. 1, Dec. 2013.

[27] N. W. Paton and A. Fernandes, “Crowdsourcing Feedback for Pay-As-You-Go Data Integration,” *DBCrowd 2013*, pp. 32–37, 2013.

[28] M. Franklin, A. Halevy, and D. Maier, “From databases to dataspaces: a new abstraction for information management,” *ACM SIGMOD Record*, vol. 34, no. 4, pp. 27–33, Dec. 2005.

[29] C. Bizer, T. Heath, and T. Berners-Lee, “Linked data-the story so far,” *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pp. 205–227, 2009.

[30] J. Hebeler, M. Fisher, R. Blace, and A. Perez-Lopez, *Semantic Web Programming*. John Wiley & Sons, 2011.

[31] Y. Lu, H. Panetto, Y. Ni, and X. Gu, “Ontology alignment for networked enterprise information system interoperability in supply chain environment,” *International Journal of Computer Integrated Manufacturing*, vol. 26, no. 1, pp. 140–151, Jan. 2013.

[32] P. Aiken and M. M. Gorman, *The Case for the Chief Data Officer*. Elsevier Inc., 2013.

[33] I. Robinson, J. Webber, J. Webber, and E. Eifrem, *Graph Databases*. Oreilly & Associates Incorporated, 2013.

[34] C. Tellkamp, A. Angerer, E. Fleisch, and D. Corsten, “From pallet to shelf: Improving data quality in retail supply chains using RFID,” vol. 17, no. 9, pp. 19–24, 2004.

[35] A. Reid and M. Catterall, “Invisible data quality issues in a CRM implementation,” *J Database Mark Cust Strategy Manag*, vol. 12, no. 4, pp. 305–314, Jul. 2005.

[36] G. Shankaranarayan, M. Ziad, and R. Y. Wang, “Managing Data Quality in Dynamic Decision Environments: An Information Product Approach,” *Journal of Database Management (JDM)*, vol. 14, no. 4, pp. 14–32, 2003.

[37] H. Nohr, “Management der Informationsqualität,” in *Informationswirtschaft*, no. 4, Wiesbaden: Deutscher Universitätsverlag, 2001, pp. 57–77.

[38] Dreibelbis, *Enterprise Master Data Management: An Soa Approach To Managing Core Information*. Pearson Education India, 2008.

[39] K. M. Hüner, M. Ofner, and B. Otto, *Towards a maturity model for corporate data quality management*. New York, New York, USA: ACM, 2009, pp. 231–238.

[40] W. W. Eckerson, “Data quality and the bottom line,” *TDWI Report*, 2002.

[41] L. P. English, *Information Quality Applied: Best Practices for Improving Business Information, Processes and Systems*. Wiley Publishing, 2009.

[42] J. E. Olson, *Data Quality*. Morgan Kaufmann, 2003.

[43] C. Batini and M. Scannapieca, *Data Quality Concepts, Methodologies and Techniques*. Springer-Verlag, 2006.

[44] T. Friedman, D. Feinberg, M. A. Beyer, and B. Gassman, “Friedman: Hype Cycle for Data Management,” *GartnerGroup Research*, 2006.

[45] R. Marsh, “Drowning in dirty data? It's time to sink or swim: A four-stage methodology for total data quality management,” *J Database Mark Cust Strategy Manag*, vol. 12, no. 2, pp. 105–112, Jan. 2005.

[46] H. A. Smith and J. D. McKeen, “Developments in practice XXX: master data management: salvation or snake oil?,” *Communications of the Association for Information Systems*, vol. 23, no. 4, pp. 63–72, Jul. 2008.

[47] P. Hopwood, “Data Governance: One Size Does Not Fit All,” *Data Governance: One Size Does Not Fit All*, 20-May-2008. [Online]. Available: http://www.information-management.com/issues/2007_48/10001356-1.html. [Accessed: 16-Jul-2014].

[48] J. R. Hudicka, *Why ETL and Data Migration Projects Fail*. Oracle Developers Technical User Group Journal, 2005.

[49] A. Lucas, “Corporate Data Quality Management in Context,” presented at the Proceedings of the th International Conference on Information Quality, 2010.

[50] E. M. Power and R. L. Trope, “Sailing in Dangerous Waters,” *American Bar Association*, 2005.

[51] B. Otto, D. Gizanis, H. Österle, and G. Danner, “Turning information and data quality into sustainable business value,” pp. 1–47, Feb. 2013.

[52] S. E. Madnick, R. Y. Wang, Y. W. Lee, and H. Zhu, “Overview and Framework for Data and Information Quality Research,” *Journal of Data and Information Quality (JDIQ*, vol. 1, no. 1, Jun. 2009.

[53] J. Vayghan, S. Garfinkle, and C. Walenta, “The IBM Data Governance Council Maturity Model: Building a roadmap for effective data governance,” *IBM Systems*, Jan. 2007.

[54] “Data Governance Part II: Maturity Models – A Path to Progress,” *NASCIO on Data Governance*, pp. 1–30, Mar. 2009.

[55] M. C. Paulk, B. Curtis, M. B. Chrissis, and C. V. Weber, “Capability maturity model, version 1.1,” *Software, IEEE*, vol. 10, no. 4, pp. 18–27, Jul. 1993.

[56] B. Parker, “Enterprise Data Management Process Maturity,” *Handbook of Data Management*, pp. 824–843, 1999.

[57] Dataflux Corporation, “The Data Governance Maturity Model. A DataFlux white paper.,” Jun. 2007.

[58] D. Newman and D. Logan, “Gartner Introduces the EIM Maturity Model,” G00160425, Dec. 2008.

[59] O. C. Helen Sun, “Enterprise Information Management: Best Practices in Data Governance,” pp. 1–20, May 2011.

[60] P. Cudre-Mauroux, *Will Graph Data Management Techniques Contribute to the Successful Large-Scale Deployment of Semantic Web Technologies?* IEEE, 2002, pp. 213–215.

[61] J. Gil-Garcia, T. Pardo, and G. Burke, “Conceptualizing information integration in government,” *Advances of Management Information Systems*, no. 17, p. 179, 2010.

[62] J. R. Gil-Garcia, C. A. Schneider, T. A. Pardo, and A. M. Cresswell, “Interorganizational Information Integration in the Criminal Justice Enterprise: Preliminary Lessons from State and County Initiatives,” *38th Annual Hawaii International Conference on System Sciences*, pp. 118c–118c, 2005.

[63] L. T. Moss, *Critical success factors for master data management*. Cutter IT Journal, 2007.

[64] L. Fernandes and M. OConnor, “Governance Should Lead the Healthcare Data Dance,” *Beye Network - Global coverage of the business intelligence ecosystem*, 20-Apr-2009. [Online]. Available: http://www.b-eye-network.com/view/10067. [Accessed: Mar-2015].

[65] R. G. J. Little and M. L. Gibson, “Perceived influences on implementing data warehousing,” *Software Engineering, IEEE Transactions on*, vol. 29, no. 4, pp. 290–296, 2003.

[66] A. Weller, “Data governance: Supporting datacentric risk management,” *Journal Of Securities Operations & Custody*, vol. 1, no. 3, pp. 250–262, 2008.

[67] J. DYCHE and K. NEVALA, “Ten Mistakes to Avoid When Launching Your Data Governance Program,” pp. 1–16, Jun. 2013.

[68] S. Nunn, “Data Governance: Make It a Priority,” *For The Record*, vol. 20, no. 19, p. 16.

[69] T. Friedman and A. Bitterer, “Magic quadrant for data quality tools,” *Gartner Group*, 2013.

[70] “Linked Data,” *w.org/standards/semanticweb/data*. [Online]. Available: http://www.w3.org/standards/semanticweb/data. [Accessed: 05-Oct-2015].

[71] T. Berners-Lee, “Linked Data,” *W3C*, 27-Jul-2006. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html. [Accessed: 10-Feb-2015].

[72] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee, *Linked data on the web (LDOW2008)*. New York, New York, USA: ACM, 2008, pp. 1265–1266.

[73] M. Manning, A. Aggarwal, K. Gao, and G. Tucker-Kellogg, “Scaling the walls of discovery: using semantic metadata for integrative problem solving.,” *Brief Bioinform*, vol. 10, no. 2, pp. 164–176, Mar. 2009.

[74] H. Alani, W. Hall, K. O'Hara, N. Shadbolt, M. Szomszor, and P. Chandler, “Building a Pragmatic Semantic Web,” *Intelligent Systems, IEEE*, vol. 23, no. 3, pp. 61–68, 2008.

[75] D. Wood, M. Zaidman, and L. Ruth, *Linked Data*. Manning Publications, 2014.

[76] T. Berners-Lee, R. Fielding, and L. Masinter, *RFC 3986: Uniform resource identifier (uri): Generic syntax*. The Internet Society, 2005.

[77] “OWL 2 Web Ontology LanguageStructural Specification and Functional-Style Syntax,” pp. 1–67, Jun. 2013.

[78] “Webster's Dictionary,” *Webster's Dictionary*. [Online]. Available: http://www.merriam-

webster.com/dictionary/ontology. [Accessed: 04-Apr-2015].

[79] G. Antoniou, E. Franconi, and F. van Harmelen, “Introduction to semantic web ontology languages,” presented at the Proceedings of the First international conference on Reasoning Web, Berlin, Heidelberg, 2005, vol. 3564, no. 1, pp. 1–21.

[80] N. F. Noy and D. L. McGuinness, “Ontology development 101: A guide to creating your first ontology,” 2001.

[81] M. PATEL and M. TRIKHA, “Interpreting Inference Engine for Semantic Web,” *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 2, no. 2, p. 676, 2013.

[82] L. D. Jan, *Enterprise ontology*. Springer, 2006.

[83] P. Groth, “Transparency and Reliability in the Data Supply Chain,” *Internet Computing, IEEE*, vol. 17, no. 2, pp. 69–71, 2013.

[84] Y. Lu, H. Panetto, Y. Ni, and X. Gu, “Ontology alignment for networked enterprise information system interoperability in supply chain environment,” *International Journal of Computer Integrated Manufacturing*, vol. 26, no. 1, pp. 140–151, Oct. 2012.

[85] A. Haller, J. Gontarczyk, and P. Kotinurmi, *Towards a complete SCM ontology: the case of ontologising RosettaNet*. New York, New York, USA: ACM, 2008, pp. 1467–1473.

[86] R. Y. Wang, “A product perspective on total data quality management,” *Communications of the ACM*, vol. 41, no. 2, pp. 58–65, Feb. 1998.

[87] A. V. Levitin, “Data as a resource: properties, implications, and prescriptions,” *Sloan management review*, pp. 89–101, 1998.

[88] Y. W. Lee and D. M. Strong, *Lee: Knowing-why about data processes and data quality - Google Scholar*. Journal of Management Information Systems, 2003.

[89] A. Haug and J. Stentoft Arlbjørn, “Barriers to master data quality,” *Journal of Enterprise Information Management*, vol. 24, no. 3, pp. 288–303, Apr. 2011.

[90] T. C. Redman, “Data Quality Management Past, Present, and Future: Towards a Management System for Data,” in *Handbook of Data Quality*, no. 2, S. Sadiq, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 15–40.

[91] T. C. Redman, *Data Driven*. Harvard Business Press, 2013.

[92] M. Villar, “Effective Business Data Stewards,” *Business Intelligence Journal*, vol. 14, no. 2, pp. 23–29, 2009.

[93] L. Alquier, T. Schultz, and S. Stephens, “Exploration of a data landscape using a collaborative linked data framework,” presented at the Proceedings of the Workshop on the Future of the Web for Collaborative Science (WWW2010), 2010.

[94] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz, “Pellet: A practical OWL-DL reasoner,” *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 5, no. 2, pp. 51–53, Jun. 2007.

[95] Y.-G. Ha, J.-C. Sohn, and Y.-J. Cho, “OWLer: a semantic web ontology inference engine,” presented at the Advanced Communication Technology, 2005, ICACT 2005. The 7th International Conference, 2005, vol. 2, pp. 1077–1080.

[96] Y. Zou, T. Finin, and H. Chen, “F-OWL: An Inference Engine for Semantic Web,” in *Formal Approaches to Agent-Based Systems*, vol. 3228, no. 16, Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 238–248.

[97] M. Jang and J.-C. Sohn, “Bossam: An Extended Rule Engine for OWL Inferencing,” in *Rules and Rule Markup Languages for the Semantic Web*, vol. 3323, no. 10, Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 128–138.

[98] J. Wesley, P. Venkatesh, and N. Kulkarni, “Semantic Wikis: Are They for You?,” *SETLabs Briefings*, vol. 7, no. 2, pp. 17–28, 2009.

[99] M. Buffa, F. Gandon, G. Ereteo, P. Sander, and C. Faron, “SweetWiki: A semantic wiki,” *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, no. 1, pp. 84–97, Feb. 2008.

[100] *Oracle Business Intelligence*. [Online]. Available: https://www.oracle.com/solutions/business-analytics/business-intelligence/index.html. [Accessed: 15-Sep-2015].

[101] “Microstrategy,” *Microstrategy Platform*. [Online]. Available: microstrategy.com/us/platform/tools. [Accessed: 15-Sep-2015].

[102] A. Fern, es, K. Belhajjame, L. Mao, and C. Guo, “A functional model for dataspace management

systems," *Advanced Query Processing*, pp. 305–341, Jan. 2013.

[103] "Protege Web Site." [Online]. Available: http://protege.stanford.edu/. [Accessed: 06-Sep-2015].

[104] "Franz Inc. Website." [Online]. Available: http://franz.com/. [Accessed: 09-Sep-2015].

[105] "W3C ConverterToRdf." [Online]. Available: http://www.w3.org/wiki/ConverterToRdf. [Accessed: 05-Oct-2015].

[106] "Apache Jena," *jena.apache.org*, 30-May-2013. [Online]. Available: https://jena.apache.org/. [Accessed: 15-Sep-2015].

[107] "Callimachus Project." [Online]. Available: http://callimachusproject.org/. [Accessed: 09-Sep-2015].

[108] M. Franklin, A. Halevy, and D. Maier, "A first tutorial on dataspaces," presented at the Proceedings of the VLDB Endowment, 2008, vol. 1, no. 2.