AnalyticsWeek article – Improving Big Data Governance with Semantics

Effective data governance consists of protocols, practices, and the people necessary for implementation to ensure trustworthy, consistent data. Its yields include regulatory compliance, improved data quality, and data's increased valuation as a monetary asset that organizations can bank on.

Nonetheless, these aspects of governance would be impossible without what is arguably its most important component: the common terminologies and definitions that are sustainable throughout an entire organization, and which comprise the foundation for the aforementioned policy and governance outcomes.

When intrinsically related to the technologies used to implement governance protocols, terminology systems (containing vocabularies and taxonomies) can unify terms and definitions at a granular level. The result is a greatly increased ability to tackle the most pervasive challenges associated with big data governance including recurring issues with unstructured and semi-structured data, integration efforts (such as mergers and acquisitions), and regulatory compliance.

A Realistic Approach

Designating the common terms and definitions that are the rudiments of governance varies according to organization, business units, and specific objectives for data management. Creating policy from them and embedding them in technology that can achieve governance goals is perhaps most expediently and sustainably facilitated by semantic technologies, which are playing an increasingly pivotal role in the overall

implementation of data governance in the wake of big data's emergence.

Once organizations adopt a glossary of terminology and definitions, they can then determine rules about terms based on their relationships to one another via taxonomies. Taxonomies are useful for disambiguation purposes and can clarify preferred labels—among any number of synonyms—for different terms in accordance to governance conventions. These definitions and taxonomies form the basis for automated terminology systems that label data according to governance standards via inputs and outputs. Ingested data adheres to terminology conventions and is stored according to preferred labels. Data captured prior to the implementation of such a system can still be queried according to the system's standards.

Linking Terminology Systems: Endless Possibilities

The possibilities that such terminology systems produce (especially for unstructured and semi-structured big data) are virtually limitless, particularly with the linking capabilities of semantic technologies. In the medical field, a hand written note hastily scribbled by a doctor can be readily transcribed by the terminology system in accordance to governance policy with preferred terms, effectively giving structure to unstructured data. Moreover, it can be linked to billing coding systems per business functions. That structured data can then be stored in a knowledge repository and queried along with other data, adding to the comprehensive integration and accumulation of data that gives big data its value.

Focusing on common definitions and linking terminology systems enables organizations to leverage business intelligence and analytics on different databases across business units. This method is also critical for determining customer disambiguation, a frequently occurring problem across vertical industries. In finance, it is possible for institutions with numerous subsidiaries and acquisitions (such as Citigroup,

Citibank, Citi Bike, etc.) to determine which subsidiary actually spent how much money with the parent company and additional internal, data-sensitive problems by using a common repository. Also, linking the different terminology repositories for these distinct yet related entities can achieve the same objective.

The primary way in which semantics addresses linking between terminology systems is by ensuring that those systems are utilizing the same words and definitions for the commonality of meaning required for successful linking. Vocabularies and taxonomies can provide such commonality of meaning, which can be implemented with ontologies to provide a standards-based approach to disparate systems and databases.

Subsequently, all systems that utilize those vocabularies and ontologies can be linked. In finance, the Financial Industry Business Ontology (FIBO) is being developed to grant "data harmonization and…the unambiguous sharing of meaning across different repositories." The life sciences industry is similarly working on industry wide standards so that numerous databases can be made available to all within this industry, while still restricting access to internal drug discovery processes according to organization.

Regulatory Compliance and Ontologies

In terms of regulatory compliance, organizations are much more flexible and celeritous to account for new requirements when data throughout disparate systems and databases are linked and commonly shared—requiring just a single update as opposed to numerous time consuming updates in multiple places. Issues of regulatory compliance are also assuaged in a semantic environment through the use of ontological models, which provide the schema that can create a model specifically in adherence to regulatory requirements.

Organizations can use ontologies to describe such requirements, then write rules for them that both restrict and

permit access and usage according to regulations. Although ontological models can also be created for any other sort of requirements pertaining to governance (metadata, reference data, etc.) it is somewhat idealistic to attempt to account for all facets of governance implementation via such models. The more thorough approach is to do so with terminology systems and supplement them accordingly with ontological models.

Terminologies First

The true value in utilizing a semantic approach to big data governance that focuses on terminology systems, their requisite taxonomies, and vocabularies pertains to the fact that this method is effective for governing unstructured data. Regardless of what particular schema (or lack thereof) is available, organizations can get their data to adhere to governance protocols by focusing on the terms, definitions, and relationships between them. Conversely, ontological models have a demonstrated efficacy with structured data. Given the fact that the majority of new data created is unstructured, the best means of wrapping effective governance policies and practices around them is through leveraging these terminology systems and semantic approaches that consistently achieve governance outcomes.

About the Author: *Dr. Jans Aasman Ph.d is the CEO of Franz Inc., an early innovator in Artificial Intelligence and leading supplier of Semantic Graph Database technology. Dr. Aasman's previous experience and educational background include:*

- *Experimental and cognitive psychology at the University of Groningen, specialization: Psychophysiology, Cognitive Psychology.*
- *Tenured Professor in Industrial Design at the Technical University of Delft. Title of the chair: Informational Ergonomics of Telematics and Intelligent Products*
- *KPN Research, the research lab of the major Dutch*

telecommunication company

• *Carnegie Mellon University. Visiting Scientist at the Computer Science Department of Prof. Dr. Allan Newell*