How a Bill Becomes a Bit: Engineering Compliance

DAVID J. MARCOS, National Security Agency

The views and opinions expressed herein are my own and do not imply endorsement by the National Security Agency or any other United States government entity.

Key Words and Phrases: compliance, privacy, information governance

1. INTRODUCTION

The Information Age presents a twofold challenge for modern organizations: complying with an ever-burgeoning set of laws, policies, and rules while simultaneously managing the torrent of ever-increasing amounts of data in the Cloud. Big Data in a highly regulated environment becomes cumbersome, requiring strict management of governing rules coupled with the ability to dynamically update and apply rules against large datasets [Breaux and Powers 2009]. However, it is possible to enumerate and comprehend multifarious rules while providing reasonable assurance that such rules are effectively applied across the breadth of Big Data. In other words, we can enable Big Compliance for Big Data with Big Rules [DeLong 2014]. The following outlines a proposed methodology for compliance engineering and an architecture to manage rules and data tagging within a highly regulated environment to act - manually or automatically - on the rules governing individual data objects. As an outline, the following should be understood as the framework wherein future papers detail specific aspects of the methodology and architecture described.

As business becomes more automated, companies must convert manual processes into automated systems. This poses several challenges, especially when dealing with large amounts of data. Principally, law and policy are human-driven processes. As such, information and process is often tacit: people intuitively understand rules; people intuitively socialize activities with one another; critically, people communicate rules verbally to one another. Attempting to take this tacit knowledge and model it explicitly into technology tends to reveal gaps that can flummox lawyers, expose procedural weaknesses, and generally, call into question an entire compliance program.

Faced with these challenges, one of three things often occurs. One, a company may simply run away from even attempting governance of its data. This approach is unlikely to succeed - particularly given the disruptive nature of modern information technology in the modern regulatory state. Two, a company may attempt to augment manual processes to discern every permutation of the rules, processes, and associated gaps that might exist. Even if this could possibly work for a time, the effort simply cannot scale with Big Data or accompanying rule sets in heavily regulated industries. Worse still, this approach engenders a false sense of security: users and automated systems may be behaving poorly - on massive amounts of information - yet such an approach may ensconce rules in process and inadvertently delay discovery of a problem. Three, a company may institute a compliance program, developing internal controls to both detect and prevent incidents. This latter option, while clearly the best option, still encounters challenges when attempting to enforce compliant behavior on data-driven business processes within the Cloud.

Managing Big Data - and managing it with myriad rules in the Cloud - is a premier challenge of the modern business world. How does one understand the totality of rules that apply to their data? How does one do so with Big Data in the Cloud? Fundamentally, how does one augment existing manual processes within a digitized environment?

For companies to be successful in the future, these questions must be addressed. Moreover, it is necessary to remove the veil that often accompanies the legalese of laws and policies: rules must be understandable to laymen (e.g., businessmen and supporting IT partners). Conversely, the implementation of the rules must also be exoteric for lawyers. Only in such an environment is it possible to comprehensively and corporately understand both data and rules, provide reasonable assurance that data is being effectively handled in accordance with applicable rules, and enable both users and systems to react to rules changes in a dynamic, facile manner [Datta, et al. 2014; DeLong 2014].

2. COMPLIANCE ENGINEERING: A NEW APPROACH

Modern companies are drowning in data and for heavily regulated industries, they are drowning in rules, too. A breath of fresh air is possible through an approach grounded in *Compliance Engineering*, i.e., the architecting of a company's IT infrastructure to support active and dynamic compliance functions against the breadth of Big Data. Fundamentally, this approach is predicated upon a data-driven business model.

This approach requires two major components. One is Rules Management: a suite of capabilities designed to digitize and manage the rules that govern business activities. The second is data tagging, specifically, Privacy and Policy-Based Data Tagging, which ties the rules as established via Rules Management to the actual data driving a business's operations.

Through this methodology, a company can improve transparency across legal/policy and operational/technological domains and see not just how a bill, regulation, or policy becomes a bit, but actively and intuitively react to new and changing rules – and data!

3. RULES MANAGEMENT

Rules Management is the set of processes and systems designed to manage the laws, policies, and rules that govern a business's operations. Many companies have such processes and systems, but they may not be actively connected to day-to-day operations. At first blush, most assume such processes and systems are primarily for reference, akin to a "library" of documentation underpinning operations used by attorneys, policymakers, and managers. But the implementation of the rules documented in these systems is often verbally communicated or rewritten into requirements for users and operational systems, risking misinterpretation that may lead to improper implementation of rules [Governatori, Milosevic and Sadiq 2006].

Rules Management, as described herein, approaches the problem differently. A "library" of rules is still necessary, but the critical step is to *connect* the rules, as documented, directly to the data that is driving operations [Datta, et al. 2014; Kagal, Hanson and Weitzner 2008]. This requires, at a minimum, three systems. These systems are distinct from one another, as described below.

3.1 Document Management

Data-driven businesses need good document management. A key step is to establish an enterprise Document Management System that provides a host of capabilities.

Foremost, the system must store the *original* documents governing operations. Additionally, these documents must be searchable in support of basic information governance and eDiscovery needs. Further, the system must support referential association and version control, allowing users to navigate through the corpus of documents, their revision histories, bibliographies, and citations. The system must also provide additional tagging capabilities to group like documents for given aspects of a business's operations, improved search-ability, and tie data-driven operations to authorizing documents [Maxwell, Anton and Swire 2011].

Lastly, the Document Management System must be authoritative. This means that tagging must have a certain exactitude that meets appropriate legal muster. Why does this matter? In subsequent sections, the connection between the tags in the Document Management System and data tags affixed to operational data will be further described, revealing that if tags are not reasonably accurate, the wrong rules might be applied to the wrong data. Thus, companies should cautiously approach performing automated tagging on governing documents using statistical algorithms often used for eDiscovery. Such tools, while often viewed as a "silver bullet" for information governance, are not guaranteed to accurately tag information in ways relevant to actual business operations, risking potential compliance incidents. Conversely, companies should also approach with care manual tagging of documents – particularly "crowd sourced" tagging, which will introduce colloquialisms and non-standardized tags that will alter the meaning and utility of document tags over time. Tagging of documents must be managed by a coterie of subject matter experts, responsible for maintaining documents as well as reviewing manual or automated tagging processes. Details regarding supporting processes to effectively operate the Document Management System and manage document tagging, while critical, are beyond the scope of this paper.

3.2 Privacy and Policy Facts

Along with a Document Management System, a second capability is necessary: the Policy Fact Engine. This system supports documentation of "authoritative facts," which are operationally actionable and intuitive to systems and users. The Policy Fact Engine is where the "rubber meets the road" – here, the rules documented in the Document Management System are enumerated to support their application in day-to-day operations.

The Policy Fact Engine lists *permissions*. Permissions consist of a set of approved activities, as dictated by governing documents stored within the Document Management System. Each entry within the Policy Fact Engine provides a list of actionable facts that instruct users and systems as to what each permission allows. Facts may include, but are not limited to, retention of information, release of information to third parties, or valid periods of time during which permissions exist. Identified facts are likely to be customized to a given business – driven by both attorneys advising which rules must be applied as well as operations and technology guiding how rules fit into operational and IT infrastructures. Once the appropriate facts are identified, it is possible to automate the extraction of permissions from the Document Management System (guided by tags applied within the Document

Management System) and auto-populate the Policy Fact Engine with apposite entries as rules appear and expire.

Returning to the idea that the Policy Fact Engine is where the "rubber meets the road," it is critical that a business identify the general facts needed to apply rules to their data. This is often a challenging ontology-modeling exercise where attorneys, operators, and technologists discuss differing perspectives of the same domain. A company must balance the need to accurately execute rules governing operations without forcing employees to become legal experts. Simultaneously, a company will need to discourage the inclination of operators and technologists to complicate the legal space through the desire to represent their activities in ways understandable to them, but that confuse the actual rules that are governing operations. The guiding principles to avoid this "metaphysical nightmare" and achieve "ontological reconciliation" should be grounded in two questions: what are the basic business functions that drive operations and what are the basic technological functions that support those business functions [Noy and McGuinness 2001]? If a company can understand and document these functions, they can reinforce their compliance architecture, grounded in their actual business practices, while also precluding "death by a thousand fact patterns." Essentially, the outcome will result in a data-driven business model through which authorities can be represented in an intuitive, lawful, reusable, and actionable manner [Shaheed, Yip and Cunningham 2005; Despres and Szulman 2004; Visser and Bench-Capon 1998].

3.3 Automated Rules Execution

The final component of Rules Management is the Automated Rules Execution Engine. This system is specifically designed to support real-time execution of given procedures as extracted from governing documentation. Like the Policy Fact Engine, this system references documents in the Document Management System that underpin the automated rules within the Automated Rules Execution Engine. Unlike the Policy Fact Engine, this system enables users and systems to *ask questions* of the rules and interactively discern answers. Whereas the Policy Fact Engine lists actionable facts *about* a given permission, the Automated Rules Execution Engine can permute over facts from *multiple* permissions and instruct users and systems as to what combinations of rules are allowable [Datta, et al. 2014; Kagal, Hanson and Weitzner 2008]. For instance, if a company has two data sets maintained under two distinct rule sets, the Policy Fact Engine will define the "ground rules" for each data set, and the Automated Rules Execution Engine will instruct as to what valid combinations of rules are allowed when combining the two data sets (such as if a user can query both data sets simultaneously or only certain aspects) [Maxwell, Anton and Swire 2011; Rosati 2006]. The key benefits here are (1) the Automated Rules Execution Engine can be queried dynamically by users and systems accessing multiple data sets and (2) the Automated Rules Execution Engine can be dynamically updated as rules change, enabling immediate application of new rules as users and systems continue to access data.

To enable such dynamism, the rules encoded in the Automated Rules Execution Engine require a high degree of scrutiny from attorneys. Fundamentally, this system is not a "lawyer in a box" that makes judgments. Rather, the system acts on pre-determined facts as dictated and approved by lawyers. Similar to the Policy Fact Engine, these "facts" must be dictated by business functions and supporting technological functions. The Automated Rules Execution Engine is acutely susceptible to "death by a thousand fact patterns." Innumerable fact patterns will

make the overall rule set inscrutable over time and profoundly inhibit performance in the Cloud. Conciseness, coupled with technological practicality, demand that automated rules be modeled in an understandable and reusable manner [Noy and McGuinness 2001].

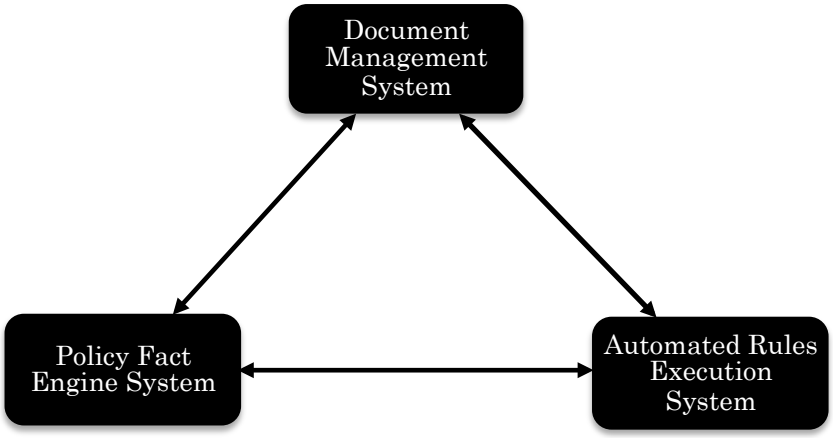

Fig. 1, Rules Management Conceptual Design (patent pending).

3.4 Why Three Rules Management Systems?

The Document Management System, Policy Fact Engine, and Automated Rules Execution Engine are distinct capabilities for two critical reasons. One is technological: the functions of each system demand very different technologies to implement. For example, combining the Document Management System and Automated Rules Execution Engine would be to the detriment of both: the Document Management System does not need a rules engine while the performance of the Automated Rules Execution Engine would be severely impacted if bound to the functionality of the Document Management System.

The second reason is sociological. The Document Management System is primarily for lawyers, policymakers, and managers in that it represents documents in ways intuitive to them, but not typically useful or understandable for operators and technologists. Thus, the Policy Fact Engine and Automated Rules Execution Engine are separate as they are strongly supportive of operational and technological representations of rules. More so, the Policy Fact Engine and Automated Rules Execution Engine serve as "mediators" between two distinct representations of the same domain: the legal/policy perspective and the operational/technological perspective. Separating the systems *forces* a necessary discussion: the proper ontological modeling of a business's functions [Noy and McGuinness 2001].

Successful modeling yields three benefits. First, it makes rules applicable and understandable to everyday employees and IT, minimizing the risk of misinterpretation. One can reliably refer to the Policy Fact Engine and Automated Rules Execution Engine to determine what to do.

Secondly, it forces attorneys to represent laws and policies in explicit, understandable terms for laymen. Focusing on specific fact patterns or deferring to

broadly applicable guidance will not work in data-driven organizations. The former will crush reusability as well as performance when working with Big Data. The latter will lead to endless interpretation by everyday employees: systems will be encoded with semi-similar rules, leading to a nightmarish test regimen to certify that systems are compliant and worse, operators and technologists will have divergent interpretations, leading operators to become unsure if systems are performing as expected [DeLong 2014]. This could lead to costly infrastructure overhauls. Meanwhile, compliance incidents will continue, impacting a company's bottom line and more importantly, its reputation with its customers.

Finally, successful modeling yields reusability: a company can start to "template" its compliance program using the model, which enables large-scale reusability in the IT space. It also reduces confusion when implemented in systems. Although difficult, it is preferable to build a "safe harbor" for developers to access rules in ways defined by company lawyers, policymakers, and managers, as opposed to instructing a developer about the rules for a data set and assuming that interpretation heard was interpretation communicated. A company will have a greater understanding of its rules as well as greater control of rule implementation and systems operations.

Rules Management, effectively implemented, becomes the "glue" that binds a business's operations and employees. It bridges perspectives, fosters reusability, and streamlines the rules governing its business model. Through Rules Management, a company can better understand itself and become significantly more responsive to its environment. Yet Rules Management is only half the battle. The key to success is *connecting* Rules Management with the actual *data* that is driving a company's operations.

4. PRIVACY AND POLICY-BASED DATA TAGGING

In the Information Age, data defines business operations. Companies must understand their data and consequently, data must be marked in such a way as to identify pertinent properties for operations. In data-driven environments companies also need to understand the rules that apply to their data – both in totality and per object [Shaheed, Yip and Cunningham 2005]. Privacy and Policy-Based data tagging is key to this understanding.

Privacy and Policy-Based data tagging covers a broad spectrum of rules that can be applied to data. A common mistake regarding data tagging is to assume governing rules can be summarized via access control [Kagal and Pato 2010]. Years ago, this may have been true – especially with stove-piped databases. In the Cloud, this approach is dangerous. Cloud intermixes different datasets, often with different rules, across myriad domains - domains that may be inapplicable or potentially illegal for a given business to use. More so, viewing rules through the prism of access control can be to the detriment of other rules. For instance, data may be governed by rules that define not just access control, but retention, usage of personally identifiable information (PII), data sharing, and other data management requirements.

There is a distinct correlation between the tags affixed to data and the modeling of business processes that yield the appropriate granularity and representation of rules stored in the Rules Management systems [Visser and Bench-Capon 1998]. Privacy and Policy-Based data tags make Rules Management a reality across a business's day-to-day operations.

Developing an effective data tagging model is not simple. Privacy and Policy-Based data tags, while directly related to the modeling of business functions, may require additional granularity to provide reasonable assurance that users and systems can effectively understand the rules that govern given data objects. Additionally, companies must be careful not to overload the meaning of Privacy and Policy-Based tags. To the greatest extent possible, data tag creation and application must be strongly governed to minimize misuse such that the tags are corporately understood and applied as defined and expected. This is similar to the need to have a group of document management subject matter experts to manage document data tagging, as noted above.

Alongside proper governance, privacy and policy-based data tagging requires accuracy. This realization is often accompanied by trepidation, as those that deal with Big Data realize the challenge of accurately tagging large amounts of data. However, it is in fact possible to develop technical controls to manage tagging accuracy, which can demonstrably reduce the risk of inaccurately applied data tags. In the near future, as more data is put into the Cloud and more regulation accompanies this data, privacy and policy-based data tagging will become increasingly necessary.

5. CONCLUSION

The Information Age is fully upon us. Technology itself is driving business. Each day, extant business processes become more data-driven. The success of future businesses depends on grappling with the implications and management of Big Data. In part, success will be determined not just by understanding the rules pertaining to a given data set, but the application of those rules against Big Data, dynamically and at scale.

The innovative methodology and architecture outlined above accomplishes this through the implementation of Rules Management coupled with Privacy and Policy-Based Data Tagging. This methodology establishes a mechanism to know the rules governing Big Data and directly connect those rules to data, bringing “rules to life” in the Cloud. This proposal calls for: (1) making rules explicit and (2) tying rules to a model reflective of a given business’s operations and people – resulting in improving compliance by making controls clear, explicit, and meaningful in day-to-day operations. Compliance Engineering, through the application of Rules Management and Privacy and Policy-Based Data Tagging, demonstrates a forward-leaning approach for businesses to manage Big Data with Big Rules.

REFERENCES

Breaux, Travis D. and Powers, Calvin. Early Studies in Acquiring Evidentiary, Reusable Business Process Models for Legal Compliance. In *6th International Conference on Information Technology: New Generations* (Las Vegas, NV 2009), IEEE, 272-266.

Datta, Anupam, Guha, Saikat, Sen, Shayak, Rajamani, Sriram K., Tsai, Janice, and Wing, Jeanette M. Bootstrapping Privacy Compliance in Big Data Systems. *IEEE Symposium on Security and Privacy* (2014).

DeLong, John. Aligning the Compasses: A Journey through Compliance and Technology. *IEEE Security and Privacy*, 12, 4 (July/August 2014).

Despres, Sylvie and Szulman, Sylvie. Construction of a Legal Ontology from a European Community Legislative Text. In *Jurix 2004: The Seventeenth Annual Conference* (Amsterdam, Netherlands 2004), IOS Press, 79-88.

Governatori, Guido, Milosevic, Zoran, and Sadiq, Shazia. Compliance checking between business processes and business contracts. In *International Enterprise Distributed Object Computing Conference (EDOC)* (Hong Kong, China 2006), IEEE, 221-232.

Guha, Saikat, Sen, Shayak, and al. *Bootstrapping Privacy Compliance in Big Data Systems*. 2014. http://www.cs.wm.edu/~ksun/csci780-f14/notes/15-Bootstrapping.ppt.

Kagal, Lalana, Hanson, Chris, and Weitzner, Daniel. Integrated Policy Explanations via Dependency Tracking. *IEEE Policy* (2008). http://dig.csail.mit.edu/2008/Papers/IEEE%20Policy/air-overview.pdf.

Kagal, Lalana and Pato, Joe. Preserving Privacy Based on Semantic Policy Tools. *IEEE Security & Privacy*, 8, 4 (Jul-Aug 2010), 25-30.

Maxwell, Jeremy C., Anton, Annie I., and Swire, Peter. *A Legal Cross-Reference Taxonomy for Identifying Conflicting Software Requirements*. Computer Science Technical Report TR-2011-4, North Carolina State University (NCSU), Raleigh, NC, 2011.

Noy, Natalya F. and McGuinness, Deborah L. *Ontology Development 101: A Guide to Creating Your First Ontology*. Knowledge Systems Laboratory Technical Report KSL-01-05, Stanford University, Stanford, CA, 2001.

Park, Jaehong and Sandhu, Ravi. The UCONABC Usage Control Model. *ACM Transactions on Information and System Security*, 0, 02 (2004).

Rosati, Ricardo. Integrating ontologies and rules: semantic and computational issues. *Lecture Notes in Computer Science (LNCS)*, 4126 (September 25-29, 2006), 128-151.

Shaheed, Jaspreet, Yip, Alexander, and Cunningham, Jim. A Top-Level Language-Biased Legal Ontology. In *International Association for Artificial Intelligence and Law (ICAIL), Workshop Series No. 4* (2005), Wolf Legal Publishers, 13-24.

Visser, Pepijn R. S. and Bench-Capon, Trevor J. M. A Comparison of Four Ontologies for the Design of Legal Knowledge Systems. *Artificial Intelligence and Law*, 6, 1 (March 1998), 27-57.