Standards-based Semantic Technologies for Smart Regulation

Professor Tom Butler

February 2017
Table of Contents

Abstract .............................................................................................................................................. 2
Introduction ......................................................................................................................................... 3
The Problems Confronting the Successful implementation of RegTech ........................................ 4
  Clarion Calls for Change .................................................................................................................. 5
Using Standards to help Solve Potential Problems with RegTech ............................................... 5
  What are Semantic Technologies? .................................................................................................. 5
Using Standards to Unpack regulatory requirements ........................................................................ 6
Using Standards-based Regulatory and Business Ontologies ....................................................... 7
  Use Case for Regulatory Ontologies .............................................................................................. 7
  Use Case for Business Ontologies .................................................................................................. 8
A Standards-based Approach to capturing Regulatory & Business Vocabularies and Rules .......... 8
Navigating the Digital Labyrinth with RegTech .................................................................................. 9
Using SemTech Standards for RegTech ............................................................................................ 11
Conclusions ....................................................................................................................................... 13

Dr Tom Butler is Professor in Business Information Systems at University College Cork and GRC Technology Centre Principal Investigator.
Abstract

This paper highlights the need for industry and technology standards in the design, implementation and use of RegTech. Without such standards, RegTech may fail to fulfil its promise of facilitating Smart regulation and become yet another dud silver bullet.

It is well-accepted that RegTech has the potential to help financial enterprises solve the regulatory interpretation problem and to meet better regulatory requirements by enabling firms to develop compliant governance and business policies. GRC solutions currently in the market are clearly deficient in this regard. Another promising RegTech feature is its ability to make regulatory compliance reporting more efficient and effective. In addition, RegTech has the ability to help firms perform better data governance and analytics, in order to support decision-making around the achievement of business objectives. It can also enable integrated risk management and to automate controls across the business. Such innovations, it is argued, will provide both firms and regulators with enhanced capabilities to detect and prevent breaches of regulatory rules. Not surprisingly, RegTech innovations also provide regulators with the ability to draft smart regulation.

This paper identifies several problems that challenge the successful design, implementation and use of RegTech. The first of these is the ‘translation problem’ and the second is the ‘Tower of Babel’ problem. As we will see, the translation problem affects not only the design and implementation of RegTech, but also how RegTech will be employed to close the gap in regulatory interpretation and understanding. The ‘Tower of Babel’ problem, however, is more fundamental—it refers to the absence of a “common language” in the financial industry. Arriving at shared business and regulatory terminological dictionaries, thesauri and taxonomies is a huge challenge—however, the lack of progress here will not only challenge RegTech, it will make the job of creating a Financial Data Standard all the more problematic.

In order to deepen our understanding of these challenges, this paper discusses how semantic standards can help solve potential problems with RegTech. Semantic technologies (SemTech) enable meaning to be attached to data—both structured and unstructured. RegTech solutions anchored on semantic standards can, for example, unpack regulatory requirements in complex and voluminous regulations. This will, we believe, require the use of standards-based regulatory and business ontologies, such as, for example, the Financial Industry Business Ontology. However, in order to ensure that this becomes a reality, the nascent RegTech sector will need to have a standards-based approach to capturing regulatory & business vocabularies and rules. Semantic standards and technologies thus developed can enable RegTech solutions help practitioners better navigate their Digital Labyrinths. SemTech will, we believe, play a key role here, as without them the challenges arising from BCBS 239 cannot be addressed in a coherent, cohesive and comprehensive manner.

Finally, the goal of this paper is to better inform the ongoing RegTech debate, and to help align RegTech initiatives with business needs, while avoiding the aforementioned problems. We identify a standards-based model that may be employed as a frame of reference. While comprehensive, it requires further elaboration and extension by the industry and its regulators. We hope that this paper will encourage a constructive dialogue in an open and collaborative manner.

This work was supported by Enterprise Ireland and the IDA under the Technology Centre Programme [Grant TC-2012-009].
Introduction

According to a recent report by Management Consultants Bain and Co, Governance Risk and Compliance (GRC) spend accounts for 15-20% of “run the bank cost”, and 40% of “change the bank costs” for major banks. Bain and Co contend that such costs will grow over the next five years, as banks continue to struggle with regulatory requirements.

There is broad agreement that banks could realize substantial benefits from innovations in RegTech in addressing this challenge. EY, for example, argue that “In the short term, adoption of RegTech will provide operational efficiencies and cost benefits when applied to current compliance and risk management practices.” The regulators appear to agree with and support the adoption of RegTech. In a speech delivered by Christopher Woolard, Director of Strategy and Competition at the Financial Conduct Authority (FCA), at London FinTech Week in July 2016, several use cases for RegTech were identified viz.

1. “First, making the business of complying with reporting requirements simpler — technology that allows more efficient methods of sharing information (for example: alternative reporting mechanisms, shared utilities and online platforms).

2. Second, technology that drives efficiencies in regulatory compliance by seeking to close the gap between the intention of regulatory requirements and the subsequent interpretation and implementation within firms. For example, we have seen a range of semantic technologies and significant enthusiasm for robo-advice style models to help firms understand their regulatory responsibilities.

3. Third, technology that simplifies and assists firms in managing and exploiting their existing data, supporting better decision-making and finding those who are not playing by the rules easier. This includes new data analytics technology, real-time compliance monitoring and trade surveillance systems.

4. Finally, technologies and innovations that allow regulation and compliance processes to be delivered differently and more efficiently. Here we see significant interest in distributed ledger technologies, automated compliance systems, machine-readable regulation and expanding use of biometrics for identity verification purposes.”

The FCA’s Project Innovate incorporates TechSprint events, the focus of two of these was RegTech and in particular the theme of Unlocking Regulatory Reporting. The GRC Technology Centre and several of its industry members attended the most recent event in February 2017. While the focus was on key aspects of regulatory reporting, the themes emerging from the discussion and presentations on nascent RegTech innovations mirror those found in thought leadership pieces and in technologies currently being deployed, for examples:

— Fraud prevention and Anti-Money Laundering (AML)
— Employee and third party surveillance
— Regulatory and governance compliance and conduct risk assessment metrics.
— Predictive analytics
— Regulatory compliance and reporting support and automation

There are varying degrees of maturity and market acceptance of these technologies. While there are clear benefits to the adoption of RegTech, there is also an unacknowledged

---

4 https://www.fca.org.uk/news/speeches/london-fintech-week-2016-innovation-regtech
5 https://www.fca.org.uk/firms/project-innovate-innovation-hub/regtech
downside. This is due, primarily to the ad hoc way in which RegTech is being adopted across the industry. The key issue here is the need for standards in the design, development and implementation of RegTech\textsuperscript{6}.

**The Problems Confronting the Successful implementation of RegTech**

In his penetrating analysis of *Technologies of Compliance* Berkeley law professor, Kenneth Bamberger, states that “While these technology systems offer powerful compliance tools, they also pose real perils. They permit computer programmers to interpret legal requirements; they mask the uncertainty of the very hazards with which policy makers are concerned; they skew decision-making through an “automation bias” that privileges personal self-interest over sound judgment; and their lack of transparency thwarts oversight and accountability. These phenomena played a critical role in the recent financial crisis.”\textsuperscript{7} One of the key issues identified by Bamberger is the problem of translation, which has several dimensions.\textsuperscript{8} There are however, other problems.

In 2013, Andrew Haldane, Executive Director for Financial Stability at the Bank of England, identified what he termed the ‘Tower of Babel’ problem. He argued that the financial industry “has no common language for communicating financial information. Most financial firms have competing in-house languages, with information systems silo-ed by business line. Across firms, it is even less likely that information systems have a common mother tongue. Today, the number of global financial

\textsuperscript{9} The scale of this problem is highlighted by the fact our research indicates that a typical large international bank may have up to 70,000 information systems and over 250,000 spreadsheets. We have previously indicated a practical consequence of this problem, best illustrated through our application of the Blind Men and the Elephant parable.\textsuperscript{10} Figure 1 illustrates this from a risk perspective.

![Figure 1 Regulatory Reporting and Risk Data](image)

This figure attempts not only the siloed nature of operational, regulatory and other risk data, but it is also evident that professional silos exist in financial services organisations. As Andrew Haldene points out, people, processes and technologies within the same organisations do not share a common language. Thus, not only do existing GRC systems suffer from translation problems, they also exhibit the ‘Tower of Babel’ problem. Without standards, RegTech will simply mean a business-as-usual approach will prevail, and the desired transformations will prove elusive.

\textsuperscript{6} See the following on the need for standards in GRC: Spies, M., & Tabet, S. (2012). Emerging standards and protocols for governance, risk, and compliance management. In Handbook of research on e-business standards and protocols: Documents, data and advanced web technologies (pp. 768-790). IGI Global.


Clarion Calls for Change
In January 2013, the Basel Committee on Banking Supervision (BCBS) issued its Principles for effective risk data aggregation and risk reporting also known as BCBS 239. This came into effect for G-SIBS, or Global Systemically Important Banks, in January 2016. These new regulatory requirements are targeted at the manner in which financial institutions manage data aggregation and manage risk. Here again the need for standards is evident in that key requirements include: Harmonization of data definitions across information systems and lifecycles; enhanced governance policies and the allocation of data ownership and accountability for the quality of risk data; and to increase data quality through the accuracy, completeness, timeliness and adaptability of data infrastructures.

In November 2015, the Financial Stability Board’s (FSB) recently called for a common language or taxonomy with which to manage Conduct Risk. It is worthwhile reproducing their requirement: “The integration of conduct risk in all aspects of a firm’s business, in a manner that is consistent across the industry, requires the development of a consistent set of definitions, methods of assessment and measurement of conduct risk.” Of course what the FSB is really requesting is a standard.

The diversity of data formats and the absence of modelling and reporting standards is also of concern to the European Commission (EC). The department for financial stability and capital markets (DG FISMA) is responsible for the EC’s policies on banking and finance. In 2016 it instituted the Financial Data Standardisation Project. Specifically, it is looking to implement Financial Data Standards for: Messaging; Semantic standards for Data Dictionaries/Ontologies/ Classification; Legal and other business identifiers, specifically entities, products, and transactions; Reporting and business domain standards; and business contract standards.

Thus, we argue that RegTech solutions providers and adopting financial institutions need to be aware of the need for standards-based approaches to the above problems, if RegTech is not to become part of the problem.

Using Standards to help Solve Potential Problems with RegTech
Semantic technologies have been identified as a means to help solve enduring problems of regulatory compliance in the financial industry. The recognition that SemTech could be of benefit to the financial industry was contemporaneous with an important and generally unnoticed paradigm shift in the IT industry with the emergence of NoSQL (Not only SQL) solutions such as graph data stores. The emergence of this new paradigm has generated new possibilities for managing, mining and processing of structured and unstructured data. However, the de facto and de jure standards that developed around semantic technologies help address the various problems with RegTech.

What are Semantic Technologies?
Semantic models and related technologies enable unstructured and structured data to be endowed with meaning—something which is not possible with traditional technologies based on relational Structured Query Language (SQL) databases or web pages based on HTML. At one level, a semantic model enables human communication. At another level a semantic model enables heterogeneous data to be linked and data in siloed SQL databases to be federated and integrated. In addition, semantic technologies can make unstructured data, such as text-based documents—e.g. regulatory texts—machine readable using domain ontologies, thereby enabling information extraction into a knowledge base.

The World Wide Web Consortium or W3C is establishing recommendations that have

---

12 http://eurofiling.info/201606/Presentations/2016-06-2%2015h00%20FDS%20Project%20Presentation.pdf
become de facto standards to support machines to process data on the WWW, which includes data in databases. Using de facto standards ensures trust and enables trusted interactions between applications in computer networks. The primary use case for Semantic Web technologies is to enable developers to store data on the Web, to build vocabularies, and to write rules for handling data. There are several core technologies which are represented in Figure 1. At the bottom of the stack is Uniform Resource Identifier (URI) which is a string of characters used to identify resource in a network. Up from this is XML—the Extensible Markup Language which defines a set of rules for structuring data and documents in a human-readable and machine-readable format. The upper layers of the stack are built on top of XML. For example, RDF or the Resource Description Framework is one of the three foundational Semantic Web technologies, the other two being SPARQL and the Web Ontology Language (OWL). RDF is the data modelling language for SemTech. OWL or the Web Ontology Language is the knowledge representation language. SPARQL or the SPARQL Protocol and RDF Query Language is, as its name indicates, the query language for the Semantic Web and siloed and distributed networked systems.

An ontology expressed in OWL provides additional semantics for data models, in that knowledge of objects and their relationships is more richly expressed. Triple stores are essentially graph stores based on RDF/RDFS while more expressive than a relational data store, are less expressive than OWL. Both can be used to capture knowledge about a domain, such as operational risk.

An ontology describes a conceptual model about a problem domain, which is in effect metadata. This can also be expressed in RDF/OWL and may be persisted in the same RDF triple store as the instance data. Thus, both metadata and data can be queried.

The power of ontologies is that they enable reasoning or inferencing in RDF triple stores. The advantage is that a reasoner may infer new/additional triples or relationships—that is add new knowledge—based on the asserted knowledge or axioms about classes and instance data in the ontology.

The W3C semantic technology stack provides an ideal platform to create extensible, standards-based RegTech platforms. This paper explains how these and related technologies can be employed for standards-based risk and compliance data aggregation in an upcoming section. However, we first address the considerable challenge of unpacking regulations and related rules into human and machine-readable formats requiring additional support for end-users. Here, again, semantic standards will play a key role.

**Using Standards to Unpack regulatory requirements**

In a perfect world, legislators and regulators would publish regulations and rules in an unambiguous, easy to interpret human- and machine-readable format. However, we do not live in such a world and firms in the financial industry face a Herculean task. It is estimated...

---

14 A complete overview of the W3C Semantic Stack may be found at https://www.w3.org/Consortium/techstack-desc.html
that 50,000 regulatory texts were published by G20 members since 2009. There is an average of 45 new documents each week. The in Markets in Financial Instruments Directive, MiFID II, has recently led to over 30,000 pages of text being generated in all aspects of its implementation.

Current approaches are labour intensive and have a lot in common with the classical Greek myth of Sisyphus. Sisyphus, the King of Ephyra, was punished for his cunning and deceitfulness by the Greek God Zeus and condemned to roll an enchanted boulder up a hill. However, Zeus’ spell ensured that it rolled back down again, leading to Sisyphus repeating the task. This destined Sisyphus to an eternity of futile, fruitless, repetitive activity.

It appears that the financial industry has been so condemned, given the volume, variety, velocity and complexity of regulations drafted since 2008, and the responses being taken to deal with the problems of regulatory compliance. As with Sisyphus, organisations typically reach the top of the hill and perform regulatory change management with boulder-sized regulations such as MiFID, for example, only to have to begin the process all over again when MiFID II came along. Generally speaking, organisations appear to be starting from scratch each time they do regulatory change management, as any previous knowledge they gained in interpreting and making sense of prior regulations has not been codified and captured in an organisational knowledge base. Dealing with regulatory rules spawned by the likes of Dodd Frank involves similar trips up and down the regulatory compliance mountain, with equally problematic outcomes for knowledge acquisition and institutional learning outcomes.

Using Standards-based Regulatory and Business Ontologies

Ontologies can help legal and business practitioners make sense of a wide and complex spectrum of legislation and regulations and to provide financial services organizations, GRC and RegTech vendors, and others in the ecosystem, with the ability to (1) query legislation, regulations, and other texts in order to identify compliance imperatives; and (2) identify changes to existing legislation and regulation introduced by amendments to existing law or new law. Thus, standards-based ontologies should inform the architecture of, or be incorporated into, RegTech solutions.

Take, for example, a variety of Upper-Level ontologies may be used (accessed via URIs) to map, integrate, semantically enrich and categorise lower level concepts and help increase overall reasoning and inferencing accuracy. URIs are Universal Resource Identifiers—these are globally unique and permit data elements (objects, classes, entities, concepts, relationships, attributes) to be identified and link data from different sources and merge them with accuracy. Thus, concepts from core ontologies such as the Financial Industry Business Ontology (FIBO) can be linked with those defined ontologies used to develop RegTech solutions.

In addition, general concepts in such ontologies can be imported from taxonomies published by the International Accounting Standards Board (IASB)/International Financial Reporting Standards (IFRS) and the US Financial Accounting Standards Board (FASB) directly, if URIs are available, or indirectly imported as concepts. Concepts and elements from FIBO, IFRS, and FASB-GAAP could form the basis of the top-half of the domain specific ontologies. In this scenario, a RegTech operational or domain-specific ontology will contain core ontology concepts and relationships and firm-specific concepts and relationships. The latter may be generated using readily available technologies from the relational schemas in operational and risk data stores, Excel schemas, or objects and relationships in unstructured data such as texts.

Use Case for Regulatory Ontologies

In keeping with the objective of standards for RegTech, the GRCTC developed the Financial Industry Regulatory Ontology (FIRO), an open

---

15 https://jwg-it.eu/insight/mifid-programme-planner/
The FIRO semantic framework is composed of four modular ontologies FIRO-H (High-level), FIRO-S (Structural), FIRO-D (Domain-specific), and FIRO-Op (Operational). The FIRO-H ontology describes high-level concepts and their relationships which are applicable across the regulatory domain. This includes concepts, such as Obligation, Prohibition, Exemption or Sanction. FIRO-S ontology models the formal structure of parliamentary, legislative, regulatory and judicial documents. FIRO-D describes domain-specific concepts and their relationships.

FIRO underpins the development of a suite of RegTech applications currently under R&D. In terms of use for regulatory compliance it can achieve the following:

- Reason on rules that are exceptions to other rules because they allow a subset of the conditions forbidden by another rule
- Reason on business rules that ensure compliance with legal rules because they require a subset of the conditions required by another rule
- Classify data (e.g. transactions) as “relevant” to a certain rule (legal statement) and further distinguish between “relevant and compliant” and “relevant and in breach of” the legal statement.

As regulatory rules reference financial processes and products, there is a necessity to have a business equivalent—and here is where business natural languages come in.

Use Case for Business Ontologies

The financial industry faces system and data integration problems that are unique in nature. Business processes and transactions span multiple entities and functions and sophisticated supply chains, with several trading entities and with data being exchanged in a range of formats and message protocols. Add to this a multiplicity of systems involved in risk and compliance management, general ledger and reporting and so on.

The major problem here is that the same data is defined differently across systems, with divergent data models and database schemes. It was with this in mind that the Enterprise Data Management (EDM) Council decided to commission a semantics model and repository for security terms and definitions to help begin to address the aforementioned problems with multiple meanings of data stored in heterogeneous databases. This would then be extended into other areas. Thus, the EDM Council recognized that the major problem facing the industry was not, necessarily, the huge volumes of data, but the different meanings attributed to the real-world objects and data entities that represent them both within and across a multiplicity of organizational information systems. Hence, in order to begin to manage the mountains of data effectively, it was recognized that the first task would be to provide a common language for the industry globally—a semantic approach was, therefore, adopted in order to arrive at unambiguous concept and relationship definitions for all financial industry data. In FIBO, concepts are defined at the business level and represented in OWL. Significantly, FIBO references other standards such as FpML, FIX, ISO, MISMO, MDDL, and XBRL.

The development and application of FIBO, as indeed FIRO, has confirmed, from both business and regulatory perspectives, the relevance of SemTech.

A Standards-based Approach to capturing Regulatory & Business Vocabularies and Rules

Given the ambiguity and complexity of legal and regulatory texts, Natural Language Processing (NLP), Machine Learning (ML) and Artificial Intelligence (AI) are not yet up to the task of unpacking regulations. Hence, the lawyer or legal subject matter expert (SME) must bear the burden of responsibility.

Our ground-breaking research and development (R&D) identified a standards-based approach that helps lawyers and legal SMEs to unpack regulations into both a human-readable and machine-computable format. The core semantic technologies we identified
are based on the Object Management Group’s (OMG) Semantics of Business Vocabulary and Business Rules (SBVR) specification—this is a de facto standard. SBVR is a specification for capturing and expressing a business vocabulary (e.g. at base a taxonomy) and business rules in a business natural language. It is grounded in ISO Common Logic and expresses rules in Deontic and Alethic Logics. SBVR was designed with business SMEs in mind, not computer scientists, who use Controlled Natural Languages.

Researchers at the GRC Technology Centre build upon SBVR to permit a lawyer or SME capture regulatory semantics and rules in a regulatory natural language (RNL). We call this Mercury. This RNL is not the controlled natural language of the computer scientist. Rather the RNL is logical, clear, unambiguous, and comprehensible by a computer programmer, while representing the regulatory semantics and rules in a human readable format. It could then be employed by the computer programmer as a specification guiding the technical implementation—avoiding the translation problem. We position Mercury as a potential de facto standard and have opened it accordingly.

An SBVR-compliant semantic repository or knowledge base typically includes a Terminological Dictionary and Rulebook. The Terminological Dictionary contains the vocabulary made up of noun concepts and verb concepts but also contains definitional rules that constrain the meaning of the entries. The Rulebook is a set of regulatory requirements in the form of behavioural and constitutive rules that capture the regulatory intent of legal texts. We also adapted SBVR and extended it to enable legal experts to perform the interpretation of regulations and capture these using our Mercury Regulatory Natural Language. We refer to this extension as Mercury-SE (Structured English). This enables the smart storage of legal interpretations in a Knowledge Base. Our SBVR-based approach also makes it possible for business SMEs to draft business vocabularies and rules on the same platform.

Together the GRCTC’s Financial Industry Regulatory Ontology (FIRO), Mercury-SE and its related XML schema, Mercury-ML (HgML) are implemented in a Web-based software application prototype called Ganesha. This application is developed in Java on the server side, and Angular JS on the client side, and the latter communicates with the server through RESTful APIs, where the vocabulary and rulebook are persisted in SQL, XML, and RDF/OWL (Resource Description Framework/Web Ontology Language 2) data stores.

It is clear from a wealth of industry feedback gained from our field research and views voiced at the recent FCA TechSprint, that standards-based RegTech architectures, such as those described above, are required.

Navigating the Digital Labyrinth with RegTech

In the Myth of the Labyrinth a Minotaur lay in wait to devour his victims. Ariadne, Mistress of the Labyrinth, helped Theseus overcome the Minotaur by providing him with a sword and a ball of golden thread—the former to slay the Minotaur, the latter to navigate his way through the maze. The myth is instructive given the significant challenge that financial institutions face in navigating through digital structured and unstructured data labyrinths without an Ariadnean Golden Thread to guide them and with the Minotaur of regulatory sanctions lying in wait.

It is evident that many financial organisations are blindly and mechanically navigating their way through the digital maze due to the limitations of traditional data management tools and techniques. Organisations cannot solve the problems they created using siloed SQL technologies in a piecemeal fashion to solve business problems by applying yet more SQL-based approaches, which do nothing to semantically enrich data or provide the capabilities to dynamically link it with other siloed internal or external data.

Thus, financial enterprises continuously repeat labour intensive processes of manually curating and integrating regulatory risk and compliance data at significant cost to the
bottom line—however, the Minotaur that is BCBS 239 also awaits the unwary and unprepared.

Our research identified how financial organisations can transcend the limitations of siloed SQL data stores and repositories of unstructured data by using standard semantic and No-SQL technologies to virtualise structured data and unlock unstructured data stored in verbatim reports, text fields and documents, thereby presenting them for semantic querying, inferencing and in-depth analysis.

In many organizations, data capture and aggregation processes that integrate structured and unstructured data from multiple siloed sources are imprecise, relatively immature, and lack the exactness to perform good data governance, let alone proper data management for risk assessment. As indicated above, organisations need to navigate a complex digital labyrinth of heterogeneous structured and unstructured data to identify, extract, transform and load data into a target platform for interpretation, analysis and reporting. It is standard practice for the majority of firms in the industry to manually curate, cleanse, and reconcile data, typically using spreadsheets, prior to the creation of aggregated management and regulatory compliance reports.

The solution to the problem of the digital labyrinth is technically feasible and practically possible, although there are few players in the market providing comprehensive solutions for the financial industry. One approach that is receiving much attention is Data Virtualisation. This approach provides access to data directly from one or more disparate data sources, without physically moving the data, and presenting it in a form that makes the technical complexity transparent to the end-user. There is broad agreement across industry sectors that semantic metadata is required to make data virtualisation and other No SQL approaches work.

In commenting on extant approaches, Richard Robinson states that “What has been missing is the centralized semantical business context, and intelligent metadata usage to create a tightly coupled, but still independent and flexible, data architecture.” However, while Brian Stein and Alan Morrison of PwC argue that the “means of creating, enriching, and managing semantic metadata incrementally is
essential,” there is a general paucity of information on the creation of semantic metadata models. All this certainly provides an opportunity for RegTech companies, particularly in light of the BCBS 239 Principles for effective Risk Data Aggregation and Risk Reporting instituted by the Basel Committee on Banking Supervision. However, compliance with BCBS 239 aside, there are compelling business drivers for effective data aggregation, which provide additional opportunities for the sector.

Figure 3 presents our proposed solution. While there are many tools to help knowledge engineers create an integrated semantic metadata model, we advise a semi-automated approach that involves the business subject matter experts (SMEs) building the metadata model according to the Object Management Group’s SBVR standard. Remember, the objective here is to create a common language to express the meaning of organisational data—only then can the apparent heterogeneity of structured and unstructured data be reconciled. RegTech applications can help achieve this if they are designed to help SMEs build both business and regulatory vocabularies and rules.

We are not alone in arguing that it is the business, and not IT, that needs to take responsibility for its data and the meanings it accords to them. Thus, business needs the tools to semantically enrich its data so that it can then virtualise it and visualise it. The next step then is to transform the business meanings to a machine readable semantic data modelling language such as the Web Ontology Language—OWL/RDF. This will then form the basis of the Integrated Semantic Metadata Model through which the structured and unstructured data may be queried and an associated Risk Data Knowledge Base populated.

One of the clear benefits of such a model is that the semantic metadata model expressed in both SBVR and OWL/RDF can be linked seamlessly (using Uniform Resource Indicators-URIs) with related semantic models like the Financial Industry Business Ontology (FIBO) and any other standards-based Knowledge Base. Also unlike traditional SQL-based approaches, the model can be extended easily. In addition, adopting such an approach avoids the double whammy of the fate of Sisyphus and the danger of being lost in the Digital Labyrinth.

**Using SemTech Standards for RegTech**

Figure 3 presents a standards-based model that may be employed as a frame of reference for the development of RegTech solutions. While comprehensive, it requires further elaboration and extension by the industry and its regulators. Support for the model’s contention that semantic technology or

---

SemTech provides the necessary and sufficient conditions for RegTech to succeed comes from Mark Robinson\textsuperscript{18}, who states that “semantic models provide the underpinning of all these technologies. They facilitate the communication between databases, applications, documents and people in extracting data from one point and transferring that message to another point—in a language that can be understood...In the RegTech world, these models can, and have, produce a semantic ontology that links the words used across regulations to describe the specific classes of requirements and how they apply to a particular regulation.”

While the above RegTech article references AI, data analytics, distributed ledger technologies, and so on, the left hand side of Figure 3 was developed based on the themes that emerged from the Financial Information Management Europe (FIMA) Conference (November, 2016). It was significant, for example, to find that the industry has yet to derive benefits from data analytics, as fundamental issues of data governance have still to be resolved. Peter Serenita, Group Chief Data Officer (CDO), HSBC, pointed out that the industry had yet to go beyond CDO 1.0 (Governance) to reach CDO 2.0 (Analytics). The panel on the implications of AI, machine learning and robotics for financial data management confirmed the pivotal role that such technologies will play in the FinTech and, particularly, RegTech domains. However, Adrian Weller, Faculty Fellow, Alan Turing Institute, stated that the real benefits of AI, in terms of unsupervised learning, are still some way off. Nevertheless, it is clear that ontologies, machine learning, and natural language processing technologies are being used effectively in the RegTech space by, for examples, RegDelta, Palantir, and others.

The points being made above by Ben Szekely in relation to the application of SemTech for enhanced data analytics and risk data aggregation in the context of smart data lakes are highly relevant, as practitioners at FIMA felt that the business benefits were neither clear nor proven. The missing ingredients in this new paradigm are a semantic layer and NoSQL technologies such as Graph or Triple Stores, as indicated by Gregory Goth in his Communications of the ACM article, The Data Lake Concept Is Maturing.\textsuperscript{19}

As expected Block Chain and Distributed Ledger Technologies (DLT) figured greatly at FIMA, as it does at most business meetings and conferences. However, the implications of DLT and Smart Contracts for RegTech is receiving attention by regulators such as the Financial Conduct Authority. As with the RegTech article cited above\textsuperscript{17}, DLT and smart contracts were proposed as a potential solution for regulatory reporting.

This last point brings us to the right hand side of the model—the SemTech stack. It was accepted by participants at the FCA’s TechSprint event that regulators, lawyers and business professionals would need an intermediate format—a regulatory, legal, and business natural language to draft smart contracts. Thus, in addition to the arguments made earlier, this is further corroboration of the need for a standard specification such as the OMG’s SBVR as a basis to develop practitioner-facing controlled natural languages.

One point that needs to be made here is that whether regulation is principles or rules based, regulators need to step up and draft regulations and rules in a human and machine readable way. Hence, the upper level of the SemTech stack falls within their area of responsibility. Financial enterprises will need to map these into governance and business policies based on a business natural language.

Both OWL (Web Ontology Language) and RDF are knowledge representation languages. An ontology expressed in OWL provides additional semantics for data models and representations, in that knowledge of objects and their relationships is more richly expressed.


through, for example, axioms. An ontology describes a conceptual model of a problem domain—viewed from another perspective, it contains metadata. Ontologies expressed in OWL may be persisted in the same RDF triple store as related instance data. Thus, both metadata—the ontology—and data—instances of classes/objects—can be queried. In addition, rule languages, such as Semantic Web Rule Language (SWRL), and so on, may be employed to add expressivity to OWL models.

Graph stores based on RDF/RDFS or other languages, while more expressive than a relational data store, can be made more expressive, when augmented with ontologies written in OWL. The power of ontologies is that they enable reasoning or inferencing in RDF triple stores. The advantage here over a graph store, for example, is that a reasoner may be used to infer new/additional triples or relationships—that is add new knowledge—based on the asserted knowledge or axioms about classes and instance data in the ontology. The question here is how to get the data into an RDF triple store from SQL and other data stores, and from unstructured data in text documents or spreadsheets.

There are two approaches:

1. Structured data in relational databases and Excel spreadsheets is extracted, transformed and loaded (ETL) into an RDF triple store or graph database. Several readily available tools perform this function. Unstructured data from text or XML documents may also be semantically enriched and mapped into an RDF triple store.

2. Structured and unstructured data in relational databases and other sources may be accessed by what is known as SPARQL endpoints. Here the data stays where it is. An endpoint is a service that permits applications to query a relational database using SPARQL, the Resource Description Framework (RDF) query language. Thus, SemTech exists that can access relational databases as virtual, read-only RDF graphs. SemTechs offer the full power of RDF-based access to data in relational databases without having to replicate it into an RDF store. Thus, for many the preferred solution is not to transform the source data into RDF, but provide the answer to the target semantic query directly from the original source data.

Typically, SemTechs that field SPARQL queries, access the relevant data stores, extract the data, transform it into RDF, and the load the RDF data into an in-memory RDF triple store for semantic querying and inferencing. Note that RDF is not the only standard format supported. Notation 3 (also known as N3) is a W3C assertion and logic language that is a superset of RDF. It extends the RDF datamodel by adding formulae, variables, logical implication, functional predicates, and other features. It is being used instead of RDF for certain applications.

One of the key challenges for RegTech is to transform unstructured data into structured data. SemTech-based solutions for this are already in use in financial service organisations. NLP technologies may be used to help semantically tag and enrich content and load it into an RDF triple store for querying. SemTech that also use a combination of machine learning and domain ontologies to query texts as unstructured data are also available. Use cases for RegTech include regulatory change management, risk management, and compliance reporting. Absent regulatory participation at the production end of the Smart Regulation, RegTech has the capabilities to make regulations smart.

It is clear from this brief overview that there is a wealth of approaches that enable standards-based technologies to apply the power of SemTech to achieve the promise of RegTech.

Conclusions

It is well-accepted that traditional technologies are not up to the task of dealing with the volume, variability, and velocity of unstructured and structured regulatory
compliance and risk data. This paper highlighted the urgency for industry and technology standards for RegTech. Without comprehensive standards, RegTech may not be the silver bullet that will help financial enterprises solve the regulatory interpretation problem and enable financial enterprises to develop compliant business models, business processes and products. Standards are also vital if RegTech is to make regulatory compliance reporting more efficient and effective. Likewise, standards will play a key role if RegTech is to have the ability to help firms perform better data governance and analytics. Standards-based RegTech can bring automation to risk identification assessment and controls, and with enhanced capabilities to detect and prevent breaches of regulatory rules. Perhaps the greatest opportunity for RegTech is to enable regulators to draft smart regulation. The achievement of these goals is, however, hampered from the outset. A number of problems exist may impact on the successful adoption and use of RegTech. The first of these is the ‘translation problem.’ Evidence has been adduced to the effect that the translation problem impacts not only the development of RegTech itself, but also the manner in which RegTech is employed to close the gap in regulatory interpretation and understanding. The second problem—the ‘Tower of Babel’ issue—is more important. This refers to the lack of a ‘common language’ in the financial industry. The lack of progress in arriving at shared business and regulatory terminological dictionaries, thesauri and taxonomies will not only imperil successful RegTech initiatives, it will impede the creation of a Financial Data Standard.

This paper discussed how SemTech-based standards can address these problems. It illustrated how RegTech can be used to enable legal and financial industry experts to transform complex legislation, related regulatory rules, and other text containing principles and standards/guidelines into a regulatory natural language (RNL). The same standards can be used to develop business natural language (BNL). The use of SemTech means that both the BNL and RNL are expressed in human- and machine-readable formats.

Of course, regulators and lawyers need to leverage the power of SemTech (ontology-enabled machine learning and NLP) to become more productive. This is the basis for Smart Regulation, at least from the consumption side of the equation. Such solutions provide a standardised, scalable, systematic approach that overcomes the limitations of current ad-hoc proprietary solutions, which see financial institutions effectively ‘reinventing the wheel’ in terms of understanding regulatory imperatives and developing related governance policies, risk management strategies, and compliance reporting solutions, whenever new legislation is published or regulations applied to industry. Semantic technologies permit meaning to be embedded in data, whether it is structured or unstructured. RegTech solutions anchored on SemTech standards can facilitate the development and use of standards-based regulatory and business ontologies and their integration with industry standard taxonomies such as IASB/International Financial Reporting Standards (IFRS) and US Financial Accounting Standards Board (FASB). As daunting as this task may seem, SemTech can now enable the semi-automatic development and enrichment of both business and regulatory ontologies.

Putting it all together, a combination of SemTech and RegTech can enable regulators and practitioners to achieve the goal of Smart Regulation, so that they can be more effective and efficient in performing regulatory compliance, and accomplish all data related activities, from aggregation to analytics, in a manner that is complaint with regulations, such as BCBS 239, MiFID II, and so on, and acts as a strategic enabler.