JOURNAL

THE CAPCO INSTITUTE JOURNAL OF FINANCIAL TRANSFORMATION 04.2017 N°45

FinTech/RegTech

Towards a Standards-Based Technology Architecture for RegTech

Tom Butler

RECIPIENT OF THE APEX AWARD FOR PUBLICATION EXCELLENCE

TRANSFORMATION

CAPCO

EMPOWERING THE FINANCIAL WORLD

Pushing the pace of financial technology, we help our clients solve technology challenges for their business – whether it's capital markets in Mumbai or community banking in Macon.

We leverage knowledge and insights from our clients around the world:

20,000 clients in towns everywhere are becoming more efficient, modern and scalable.

27 billion transactions processed help solve clients' challenges – big and small.

$9 trillion moved across the globe in a single year empowers our clients' communities to build storefronts, homes and careers.

55,000 hearts and minds have joined forces to bring you greater capabilities in even the smallest places.

Empowering the Financial World

FISGLOBAL.COM

FIS

JOURNAL

THE CAPCO INSTITUTE JOURNAL OF FINANCIAL TRANSFORMATION

Recipient of the Apex Award for Publication Excellence

Transformation

FinTech/RegTech

Investments

Banking

Towards a Standards-Based Technology Architecture for RegTech[1]

Tom Butler – Professor, GRC Technology Centre, University College Cork, Ireland

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." It is well-accepted that RegTech has the potential to help financial enterprises address the following issues: (1) solve the regulatory interpretation problem; (2) develop compliant governance and business policies; (3) make regulatory compliance reporting more efficient and effective; (4) help firms perform better data governance and analytics; (5) enable integrated risk management; and (6) automate controls across the business. Two significant problems challenge the potential of RegTech. The first of these is the "translation problem," which affects not only the design and implementation of RegTech, but also how it will be employed to close the gap in regulatory interpretation and understanding. The second is the "Tower of Babel" problem, which refers to the absence of a "common language" in the financial services industry. This paper discusses how semantic standards can help solve potential problems with RegTech. Semantic technologies enable meaning to be attached to data – both structured and unstructured. RegTech solutions anchored on semantic standards can unpack regulatory requirements in complex and voluminous regulations. This will, we believe, require the use of standards-based regulatory and business ontologies. Semantic standards and technologies thus developed can enable RegTech solutions to help practitioners better navigate their digital labyrinths. Semantic technologies 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.

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

INTRODUCTION

According to a recent report by 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.[2] 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.[3] 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."[4] 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,[5] 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 has been RegTech and, in particular, the theme of "unlocking regulatory reporting."[6] 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, such as: (1) fraud prevention and anti-money laundering (AML); (2) employee and third party surveillance; (3) regulatory and governance compliance and conduct risk assessment metrics; (4) predictive analytics; and (5) 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 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.[7]

THE PROBLEMS CONFRONTING THE SUCCESSFUL IMPLEMENTATION OF REGTECH

In his penetrating analysis of "technologies of compliance," 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."[8] One of the key issues identified by Bamberger is the problem of translation, which has several dimensions.[9] There are, however, other problems.

2 http://bit.ly/2devi2n
3 Arner, D. W., J. Barberis, and R. P. Buckley, 2016, "The Emergence of RegTech 2.0: from know your customer to know your data," Journal of Financial Transformation 44, 79-86
4 https://go.ey.com/24SGCnl
5 http://bit.ly/2m2UH54
6 http://bit.ly/2ffadWC
7 See the following on the need for standards in GRC: Spies, M., and S. Tabet, 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, IGI Global.
8 Bamberger, K. A., 2009, "Technologies of compliance: risk and regulation in a digital age," Texas Law Review, 88:4, 669-739
9 Butler, T., and E. Abi-Lahoud, 2014, "A hermeneutic approach to solving the translation problem in designing ontologies," 22nd European Conference on Information Systems (ECIS), Tel Aviv, Israel

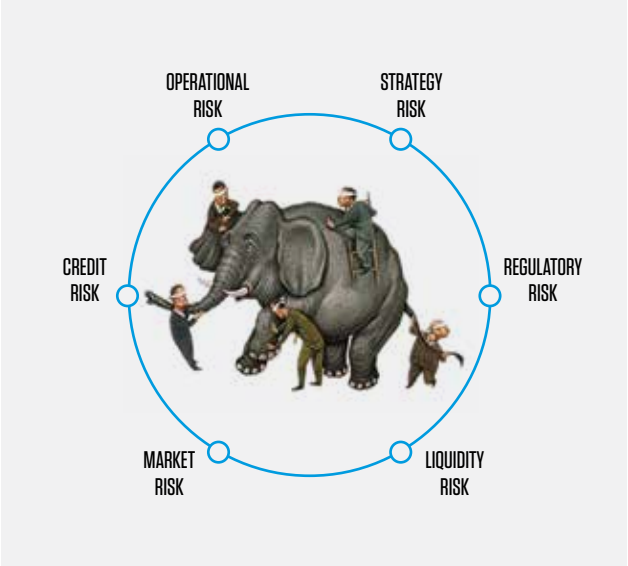

Figure 1 – Regulatory reporting and risk data

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 siloed by business line. Across firms, it is even less likely that information systems have a common mother tongue. Today, the number of global financial languages very likely exceeds the number of global spoken languages."[10] The scale of this problem is highlighted by the fact that 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.[11] Figure 1 illustrates this from a risk perspective.

Figure 1 attempts to not only highlight the siloed nature of operational, regulatory, and other risk data, but also the fact that professional silos exist in financial services organizations themselves. As Andrew Haldene points out, people, processes, and technologies within the same organizations 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 that a business-as-usual approach will prevail, and the desired transformations will prove elusive.

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 risk. Here again the need for standards is evident in that key requirements include: (a) harmonization of data definitions across information systems and lifecycles; (b) enhanced governance policies and the allocation of data ownership and accountability for the quality of risk data; and (c) improved data quality through the accuracy, completeness, timeliness, and adaptability of data infrastructures.

In November 2015, the Financial Stability Board's (FSB) called for a common language or taxonomy with which to manage conduct risk.[12] It is worthwhile restating their requirements here: "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 modeling and reporting standards is also of concern to the European Commission (E.C.). The department for financial stability and capital markets (DG FISMA) is responsible for the E.C.'s policies on banking and finance. In 2016, it instituted the Financial Data Standardization Project.[13] Specifically, this is looking to (a) implement financial data standards for messaging; (b) semantic standards for data dictionaries/ontologies/classification; (c) legal and other business identifiers, specifically entities, products, and transactions; (d) reporting and business domain standards; and (e) 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 itself.

10 Haldane, A. G., 2012, "Towards a common financial language," presentation at the Securities Industry and Financial Markets Association (SIFMA) "Building a global legal entity identifier framework" Symposium, New York, 14 March
11 Butler, T., and E. Abi Lahoud, 2014, "Applying semantic technologies for risk data aggregation," Consortium for System Risk Analytics (CSRA) meeting, MIT Sloan Center for Finance and Policy, Cambridge, MA, December
12 http://bit.ly/2m3g4Si
13 http://bit.ly/2llZFvg

USING STANDARDS TO HELP SOLVE POTENTIAL PROBLEMS WITH REGTECH

Semantic technologies (SemTech) have been identified as a means to help solve enduring problems of regulatory compliance in the financial services industry. The recognition that SemTech could be of benefit to the 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.[14] 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 sematic technologies help address the various problems with RegTech.

What is SemTech?

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, SemTech can make unstructured data, such as text-based documents, such as regulatory texts, machine readable using domain ontologies, thereby enabling information extraction into a knowledge base.

The World Wide Web Consortium (W3C) is establishing recommendations that have become de facto standards for supporting machines in processing 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 that are represented in Figure 2.[15] At the bottom of the stack is "uniform resource identifier" (URI), which is a string of characters used to identify resource in a network. Above it is XML (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 (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 modeling language for SemTech. OWL 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.

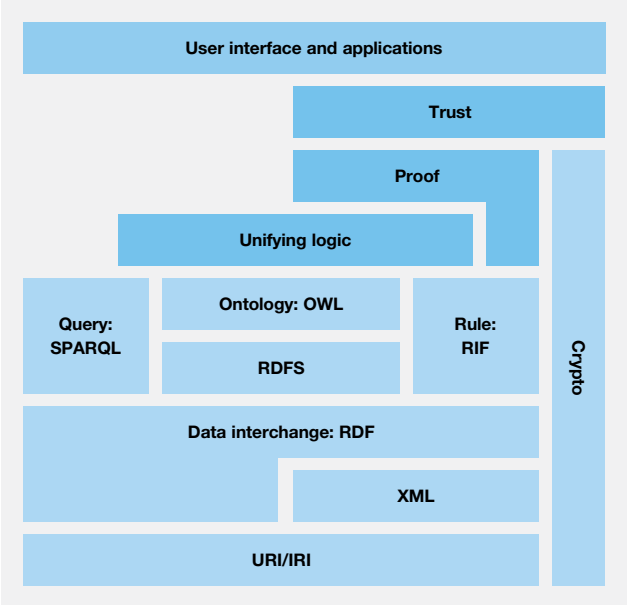

Figure 2 – W3C semantic web stack

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, and while they more expressive than a relational data store they 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.

14 McCreary, D., and A. Kelly, 2013, Making sense of NoSQL, Manning Publications, Greenwich

15 A complete overview of the W3C Semantic Stack may be found at http://bit.ly/2mAGEVT

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 services industry face a Herculean task. It is estimated that 50,000 regulatory texts were published by G20 members since 2009. There is an average of 45 new documents each week.[16] MiFID II (Markets in Financial Instruments Directive) has recently led to over 30,000 pages of text being generated in all aspects of its implementation.

Current approaches to unpack regulatory requirements are labor 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 services 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, organizations 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, organizations 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 organizational 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.

For example, a variety of upper-level ontologies may be used (i.e., accessed via URIs) to map, integrate, semantically enrich, and categorize lower level concepts and help increase overall reasoning and inferencing accuracy. URIs (universal resource identifiers) are globally unique, 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 U.S. 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.

16 http://bit.ly/2lCoQVy

USE CASE FOR REGULATORY ONTOLOGIES

In keeping with the objective of standards for RegTech, the GRCTC (Governance Risk and Compliance Technology Centre at University College Cork) developed the "financial industry regulatory ontology" (FIRO), an open standardized model of regulations. 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 services 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 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, 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 and 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 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 GRCTC 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 behavioral 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 RNL. 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 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 organizations are blindly and mechanically navigating their way through the digital maze due to the limitations of traditional data management tools and techniques. Organizations cannot solve the problems they created using siloed SQL technologies in a piecemeal fashion 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 labor-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 organizations can transcend the limitations of siloed SQL data stores and repositories of unstructured data by using standard semantic and No-SQL technologies to virtualize 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.

SemTech and risk data aggregation

"The foresight required for long-term sustainability of smart data lakes is embedded within a semantic model, which provides conceptual descriptions of data via ontologies and visually represents them, their attributes, and their relationships to other data via graph technologies. These descriptions and different data elements are useful for metadata management, mapping, and linking data as needed, and provide the foundation for ensuring governance protocols, data discovery, preparation processes, and more. The graph-based model and detailed descriptions of data elements they enable substantially enhance integration efforts, enabling business users to link data according to relevant attributes that provide pivotal context across data sources and business models. The result is considerably decreased time to a more profound form of analytics, in which users can not only ask more questions more expediently than before, but also determine relationships and data context to issue ad-hoc queries for specific needs."[17]

17 Szekely, B., 2015, "Avoiding three common pitfalls of data lakes," http://bit.ly/2mARrQ9

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, organizations 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 services industry. One approach that is receiving much attention is "data virtualization." 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 virtualization and other NoSQL 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."[18] However, while Brian Stein and Alan Morrison of PwC argue that the "means of creating, enriching, and managing semantic metadata incrementally is essential,"[19] 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. 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 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 organizational 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.

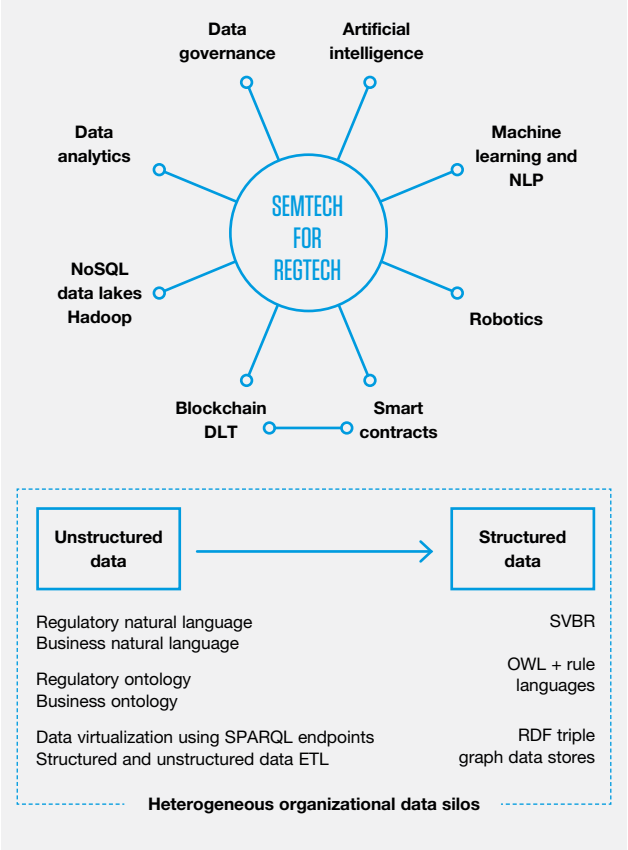

Figure 3 – SemTech for RegTech

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 virtualize it. The next step then is to transform the business meanings to a machine readable semantic data modeling language, such as 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 URIs) with related semantic models like FIBO and any other standards-based knowledge

18 http://bit.ly/2mj1g3V
19 http://pwc.to/2lmkUwT

base. In addition, unlike traditional SQL-based approaches, the model can be extended easily. Further, 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,[20] 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.[21]

As expected, blockchain 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. Similar to what Szekely stated, 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 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), may be employed to add expressivity to OWL models.

20 Robinson, M., 2016, "The RegTech marketplace: in depth analysis," http://bit.ly/2m3J26b

21 Goth, G., 2016, "The data lake concept is maturing," http://bit.ly/2mTPSc2

Graph stores based on RDF/RDFS or other languages, while more expressive than a relational data store, may become more expressive, when augmented with ontologies written in OWL, in capturing knowledge about a domain, such as fund management and related topics like regulatory or operational risk.

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 are 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 RDF query language. Thus, SemTech accesses 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 field SPARQL queries, access the relevant data stores, extract the data, transform it into RDF, and then 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 data model 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 organizations. NLP technologies may be used to help semantically tag and enrich content and load it into an RDF triple store for querying. SemTechs 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 many perceive it to be in order to help financial enterprises solve the regulatory interpretation problem and enable them to develop compliant business models, 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, however, is to enable regulators to draft smart regulation.

The achievement of these goals is, nevertheless, hampered from the outset. A number of problems exist that may have an 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 it 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 services 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 standardized, 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 the U.S. 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.

Layout, production and coordination:
Cypres – Daniel Brandt, Kris Van de Vijver and Pieter Vereertbrugghen

BANGALORE
BRATISLAVA
BRUSSELS
CHICAGO
DALLAS
DÜSSELDORF
EDINBURGH
FRANKFURT
GENEVA
HONG KONG
HOUSTON
KUALA LUMPUR
LONDON
NEW YORK
ORLANDO
PARIS
SINGAPORE
TORONTO
VIENNA
WASHINGTON D.C.
ZURICH

CAPCO.COM