activities is extrapolated from specializations presented in Section 2. We propose exactly how the (semi-)automation may be achieved for each activity in the set using model-driven techniques. Section 4 discusses how this compliance framework may lead to cost efficiency and business effectiveness. Section 5 concludes the paper.

2. RELATED AND PREVIOUS WORK

In the following, we consider regulatory requirements from external regulatory bodies as the key source, although our discussion is applicable to policies internal to an enterprise.

2.1 Generic Set of Activities in Compliance

Compliance checking can be classified based on whether it is design-/run-time depending on whether information required for checking is available only at run-time. It can also be classified as forward or backward checking based on whether controls are enacted in processes preemptively or execution traces are checked after business processes have already executed. Another way to classify compliance checking is what the granularity of checks is, i.e., whether business processes, tasks, or attributes or pure data is checked, and finally whether checking takes place by making use of an inference engine and/or queries to models of enterprise information [12]. Several works have surveyed existing compliance checking approaches from academia based on similar classification of compliance checking activities [6] and also from industry governance, risk, and compliance (GRC) approaches [10].

For the purpose of this paper, we limit the generic set of activities and artifacts in compliance to those illustrated in Figure 1. Legal text indicates the source of regulations, which could be a document from a regulatory body in a give domain or an interpretation by various stakeholders of an enterprise. The regulations and/or interpretation are predominantly natural language texts. Enterprise information against which regulations specified in legal texts are to be checked can manifest in number of forms including natural language texts, operational models including business process definitions, execution traces, or audit trails, or databases. Compliance checking and report generation involves specifying rules from legal text and facts from enterprise information in a suitable format and performing the checking activity. Note that industry GRC approaches primarily use querying mechanisms as opposed to compliance engines as in academia for checking compliance.

We showed in our earlier works that formal approaches from academia often *assume implicitly* that terms in legal texts and enterprise information artifacts match [24]. This is indicated by an optional artifact called *vocabulary* in Figure 1. Several combinations of rule and operational specifications exist in academic literature with implicit assumptions about terms in both [24]. Industry GRC approaches use *taxonomy* as the collection of predefined tags available for enterprises to *affix* to their financial data [4]. Tags can be specific to territories/geographies, time frames, and business units. Tags either do not leverage semantic meanings of terms or the support for such semantics is rudimentary at best in most GRC approaches [24].

Both kinds of approaches vary based on constituents of compliance, i.e., the legal and enterprise information artifacts, their formats, or formalisms and the purpose of compliance. In the next section, we show variations of both

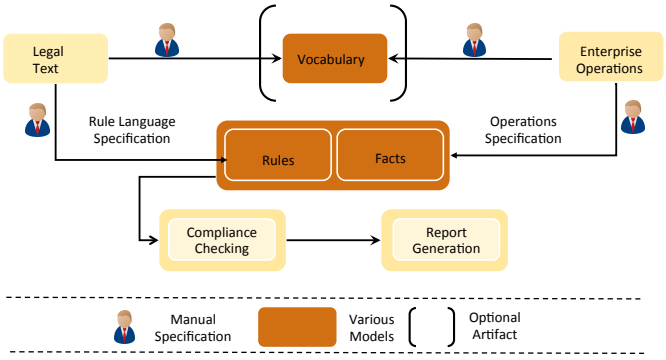

Figure 1: Generic Set of Activities and Artifacts in Compliance Management.

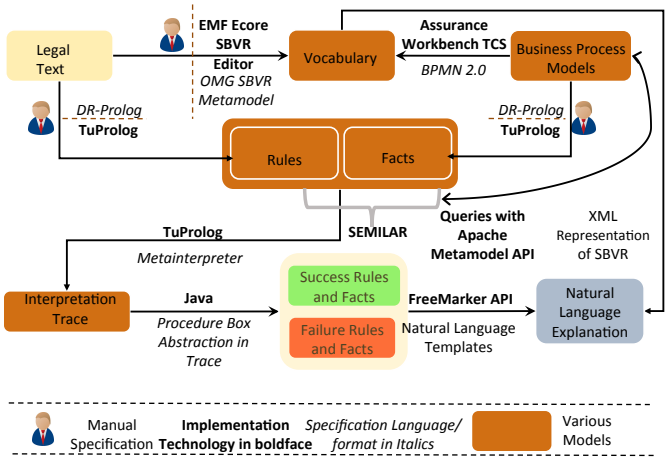

Figure 2: Explanation of Proof/Evidence of Compliance.

constituents and purposes as specialization of generic set of activities from our previous work.

2.2 Purpose and Constituents of Compliance

We show two specializations where the enterprises may have process or just the data and the purpose may be to obtain proof/evidence of (non-)compliance or to generate reports of violation based on auditors' demand.

2.2.1 *Explanation of Proof/Evidence of Compliance*

We utilized a specialization of generic set of activities in Figure 1 as illustrated in Figure 2. This was our attempt to leverage the holistic perspective of governance, risk, and compliance from industry GRC approaches along with formal treatments as in academic approaches. In this case, the constituents of compliance are legal text and business process (BP) models. While process models are BP modeling notation 2.0 compliant, we utilize DR-Prolog as the specification language for both rules obtained from legal texts and facts extracted from process models. In addition to compliance checking the specialization in Figure 2 enables natural language explanation of proofs of (non-) compliance.

We build the vocabulary model based on Semantics of Business Vocabulary and Rules (SBVR) metamodel from the SBVR specification [19]. The vocabulary model repre-

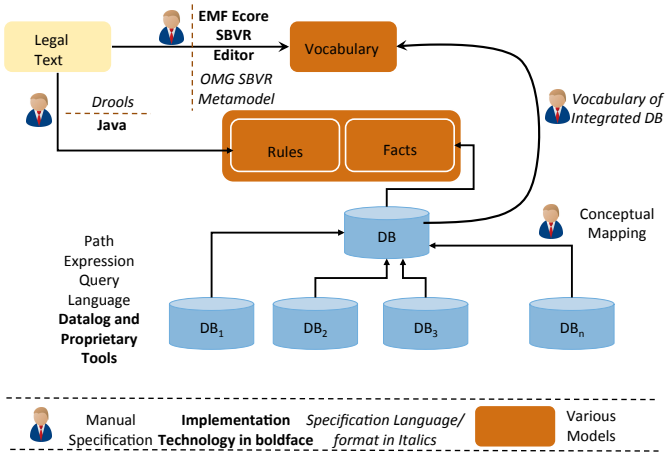

Figure 3: Compliance Report Generation using Multi-source Data.

sents terms from the legal text and BP models. These terms from legal and business side are reconciled using SEMILAR similarity measurement API [20]. We use DR-Prolog defeasible compliance engine to express rules from legal text and facts in relational form that we extract from BP models using a proprietary tool. Specialized algorithms using a Prolog-based meta-interpreter emit a suitable trace which is parsed to obtain rules and facts contributing to success or failure of queries of compliant rules. The terms from this subset of rules and facts are matched with the vocabulary to express the natural language explanation using FreeMarker template API where relevant details of terms under consideration are inserted into the variable parts of a template. We redirect the reader to [22] for details of proof generation and natural language explanation.

This specialization is useful when an enterprise aims at obtaining explanation of proofs of (non-) compliance in addition to checking whether its operational practices are compliant or not with given set of regulations. We demonstrated the utility of this framework on a real world *Know Your Customer* (KYC) regulations by Reserve Bank of India for Indian Banks [22]. In this particular case, the bank had business processes where both backward checking (checking data generated by processes) and forward checking (realizing controls on specific activities) could be achieved.

2.2.2 *Report Generation with Multi-source Data*

The second specialization that we illustrate is a work-in-progress where the enterprise has data instead of process descriptions. This data is to be obtained from sources in various business units. Figure 3 shows this use case.

The databases DB_1 to DB_n hold the data which is integrated into the database DB. We use our proprietary tooling for this purpose where a specialization of object query language called path expression query language is used for mapping conceptual models of DBs to the integrated DB. The actual model processing uses Datalog and other proprietary tools described in detail in [28]. The rule specification language used in this case is Drools which takes plain old Java objects (POJOs) as the fact model which is checked against rules implementing *Rete* pattern matching. Both the vocabularies of legal text and the integrated DB are created and reconciled in a manner similar to as detailed in our earlier work [24]. The reports are generated using Drools reporting features, but mostly contain information of checked passed and failed rather than explanations of the same.

3. (SEMI-)AUTOMATED AND END TO END COMPLIANCE

Two specializations we described in Figures 2 and 3 show that depending on the enterprises' purpose and the form of operational specifics it has, the generic set of activities in Figure 1 can be specialized. Referring back to Section 1, our specializations use models for representing rules and facts and also for expressing semantic similarity between the vocabularies of legal texts and enterprise information. More information about how we create SBVR-based models and how we utilize SEMILAR for contextual similarity measurement can be found in [24]. To some extent, this satisfies requirement (a) that of relating compliance requirements to business operations. These models together enable end-to-end compliance management in various combinations of rule and operation specifications as described in the previous section and thereby satisfies requirement (b) to a large extent. Similarly, we demonstrated in [23], how risks pertaining to the compliance of given set of regulations and corresponding mitigation activities can be modeled and how to utilize these models. This satisfies requirement (c) to some extent.

Yet, most of the activities continue be manual as evident in Figures 2 and 3, which need to be automated to the extent possible. Also, to effectively relate business objectives to compliance, further modeling and model processing machinery is required. We propose how this can be done next.

3.1 Automating Model Generation for Rules and Facts

To represent rules and facts from legal text and enterprise information, it might be possible to extract each using natural language processing (NLP) and machine learning (ML). There exists sizable literature on extracting conceptual models of regulation or rules from legal/regulatory texts. Most of these approaches focus on using either a simplified representations of natural language texts or making assumption about structural aspects of the texts or both. We review such proposals next briefly.

An approach presented in [27] uses a modeling interface that the domain expert (referred to as a *knowledge engineer* in [27]) can use to build the conceptual model and norms incrementally. At the back of this interface are a set of NLP components including a parser, a grammar, a lexicon, and a lexicon supplementor for identifying grammatical categories, all of which are specific to Dutch language. They make a suitable assumption that a set of possible *juridical natural language constructs* (JNLC) can describe categories like definitions, value assignments, and conditions. If the regulation text does not contain presumed syntactic structures then it has to be rewritten to make the syntactic structures explicit. Only when the syntactic structures are explicit that a parser written to identify them can be actually used.

A similar approach for Italian language is presented in [18] which uses articles, sections, and paragraphs to identify especially the amendments to original laws. Breaux et al. propose a systematic *manual* process [7], in which the domain expert marks the text using phrase heuristics and a

frame-based model to identify rights or obligations, associated constraints, and condition keywords including natural language conjunctions. These rights and obligations are restated into restricted natural language statements (RNLS). The RNLS can be modeled as description logic rules using semantic parameterization process. Kiyavitskaya et al. proposed to add tool support to this process [13], which was carried out in work by Zeni et al. [29]. In this work, a document structure is assumed with varying granularity from words and phrases to sections and documents. Various syntactic indicators are used to capture deontic concepts and exceptions. For instance, the concept of *right* is identified in the text via indicators like *may, can, could, permit, to have a right to, should be able to*. Some of the indicators could be complex patterns that combine literal phrases and basic concepts. The *annotation schema* that specifies rules for identifying domain concepts via indicators is nevertheless created mostly manually, whereby authors plan to use clustering techniques to automate the same.

In these approaches, domain experts are required to annotate the text initially to explicate the core concepts, syntactic structures, or patterns which are then incorporated in parsing. Domain experts may also have to rewrite the text in a simpler form for it to become amenable to specialized parsing mechanisms. The problem with these approaches is that they are very specific to a kind of regulation with parsing mechanisms specialized around the syntactic structures of that regulation. A generic set of NLP-ML techniques is more amenable than coming up with individual set of techniques for each. There are several pointers for improvement with this state of the art:

- We may take a clue from taxonomy tagging tools from industry such as OpenCalais[1], Active Tags[2], and Compliance Guardian[3] to initially present a list of important concepts from the text to the domain expert. These concepts could be top-k concepts frequency distribution-wise.
- Alternatively, domain experts may suggest a few concepts core to the regulation which can be used as seeds to obtain an initial conceptual model which can be incrementally built to include necessary and sufficient concepts.
- Instead of using regulation-specific heuristics, one could use phrase heuristics for building domain models based on identification of entities, attributes, and relations as applied to regular text. There are several works in NLP-ML, which use a variety of heuristics and training methods targeted at creating concept hierarchies via syntactic heuristics, semantic patterns, and un- and (semi-) supervised methods [2, 9, 21, 14].
- Note that most of works on legal text extraction do not consider enterprise information against which regulations are to be checked. The NLP-ML techniques need to be applied to enterprise information as well, available in the form of business process definitions, data, or audit trails, to obtain a conceptual model with which to map regulation concepts.

Once the NLP-ML techniques are applied to obtain conceptual models and a mapping between them, this model can be used to generate rules and facts in the desired specification language. We presented early manifestation of the idea of generating requisite artifact specifications from a conceptual model in [23]. At this stage, a (semi-)automated specialization of generic set of activities in Figure 1 could be imagined as illustrated in Figure 4.

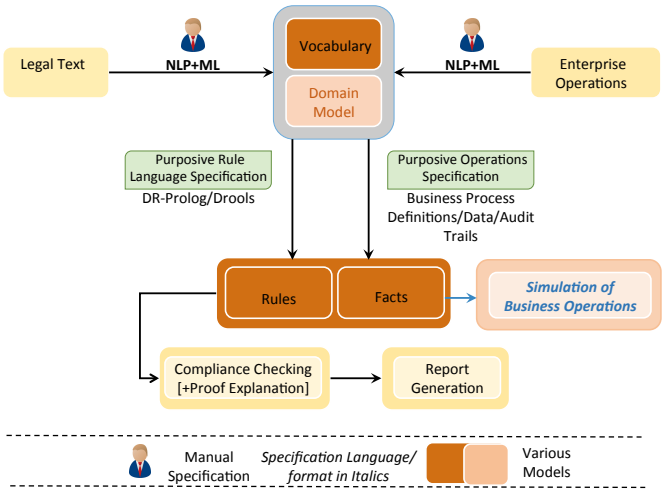

Figure 4: (Semi-) Automation with Purposive Compliance

Compared to the generic set of activities for compliance management and its specializations presented in Section 2, the framework illustrated in Figure 4, restricts the role of domain experts in conceptual model making. The process of generation of model is (semi-)automated since we envision that such a model will be built incrementally along the lines of approach presented in [27] which we reviewed earlier. This conceptual model needs to incorporate concepts of risks and governance as we indicated in [23]. With this set of concepts, it might be possible to simulate operations of enterprise to get a better fit between compliance and business objectives as described next.

3.2 Simulating Operations with Compliance Controls

According to the recent Mckinsey report on global risk practice [11], in the traditional compliance management, business managers are left to their own devices to figure out specific controls required to address regulatory requirements, leading to build up of labor-intensive control activities with uncertain effectiveness. Compliance activities tend to be isolated, lacking a clear link to the broader framework of underlying risks and business goals with a dramatic increase in compliance and control spend with either limited or unproved impact on the residual risk profile of given enterprise.

In our prior work, we modeled existing operational practices of enterprises using enterprise architecture and business motivation models [26]. We also showed how to incorporate directives such as internal policies and external regulations in enterprises' to-be architecture [25]. An enterprise needs to maintain both its *business as usual* state and to keep it optimum with regards certain criteria and it is also in-

[1]OpenCalais (Thomson Reuters) http://new.opencalais.com/opencalais-api/
[2]Active Tags http://www.wavetrend.net/activ-tags.php
[3]Compliance Guardian http://www.avepoint.com/products/compliance-management/

[9] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1535–1545. ACL, 2011.

[10] J. A. W. French Caldwell. Magic quadrant for enterprise governance, risk and compliance platforms, 2013.

[11] P. Kaminski and K. Robu. Compliance and control 2.0: Emerging best practice model. *McKinsey Working Papers on Risk*, 33, Oct 2015.

[12] M. E. Kharbili, A. K. A. de Medeiros, S. Stein, and W. M. P. van der Aalst. Business process compliance checking: Current state and future challenges. In P. Loos, M. Nüttgens, K. Turowski, and D. Werth, editors, *MobIS*, volume 141 of *LNI*, pages 107–113. GI, 2008.

[13] N. Kiyavitskaya, N. Zeni, T. D. Breaux, A. I. Antón, J. R. Cordy, L. Mich, and J. Mylopoulos. Automating the extraction of rights and obligations for regulatory compliance. In Q. Li, S. Spaccapietra, E. S. K. Yu, and A. Olivé, editors, *Conceptual Modeling - ER 2008, Barcelona, Spain*, volume 5231 of *Lecture Notes in Computer Science*, pages 154–168. Springer, 2008.

[14] I. Klapaftis. *Unsupervised concept hierarchy induction: learning the semantics of words*. University of York, Department of Computer Science, 2009.

[15] KPMG. The convergence evolution: Global survey into the integration of governance, risk, and compliance, Feb 2012.

[16] V. Kulkarni, S. Barat, T. Clark, and B. S. Barn. Toward overcoming accidental complexity in organisational decision-making. In Lethbridge et al. [17], pages 368–377.

[17] T. Lethbridge, J. Cabot, and A. Egyed, editors. *18th ACM/IEEE International Conference on Model Driven Engineering Languages and Systems, MoDELS 2015, Ottawa, ON, Canada, September 30 - October 2, 2015*. IEEE, 2015.

[18] P. Mercatali, F. Romano, L. Boschi, and E. Spinicci. Automatic translation from textual representations of laws to formal models through UML. In M. Moens and P. Spyns, editors, *Legal Knowledge and Information Systems - JURIX 2005: The Eighteenth Annual Conference on Legal Knowledge and Information Systems, Brussels, Belgium, 8-10 December 2005*, volume 134 of *Frontiers in Artificial Intelligence and Applications*, pages 71–80. IOS Press, 2005.

[19] OMG. Semantics of business vocabulary and business rules (SBVR), v1.3. May 2015.

[20] V. Rus, M. C. Lintean, R. Banjade, N. B. Niraula, and D. Stefanescu. SEMILAR: the semantic similarity toolkit. In *51st Annual Meeting of the Association for Computational Linguistics, ACL, Sofia, Bulgaria*, pages 163–168. The Association for Computer Linguistics, 2013.

[21] I. Serra, R. Girardi, and P. Novais. Evaluating techniques for learning non-taxonomic relationships of ontologies from text. *Expert Syst. Appl.*, 41(11):5201–5211, 2014.

[22] S. Sunkle, D. Kholkar, and V. Kulkarni. Explanation of proofs of regulatory (non-)compliance using semantic vocabularies. In N. Bassiliades, G. Gottlob, F. Sadri, A. Paschke, and D. Roman, editors, *Rule Technologies: Foundations, Tools, and Applications - 9th International Symposium, RuleML 2015, Berlin, Germany, August 2-5, 2015, Proceedings*, volume 9202 of *Lecture Notes in Computer Science*, pages 388–403. Springer, 2015.

[23] S. Sunkle, D. Kholkar, and V. Kulkarni. Model-driven regulatory compliance: A case study of "know your customer" regulations. In Lethbridge et al. [17], pages 436–445.

[24] S. Sunkle, D. Kholkar, and V. Kulkarni. Toward better mapping between regulations and operations of enterprises using vocabularies and semantic similarity. *CSIMQ*, 5:39–60, 2015.

[25] S. Sunkle, D. Kholkar, H. Rathod, and V. Kulkarni. Incorporating directives into enterprise TO-BE architecture. In G. Grossmann, S. Hallé, D. Karastoyanova, M. Reichert, and S. Rinderle-Ma, editors, *18th IEEE International Enterprise Distributed Object Computing Conference Workshops and Demonstrations, EDOC Workshops 2014, Ulm, Germany, September 1-2, 2014*, pages 57–66. IEEE, 2014.

[26] S. Sunkle and H. Rathod. Visual and ontological modeling and analysis support for extended enterprise models. In S. Nurcan and E. Pimenidis, editors, *Information Systems Engineering in Complex Environments - CAiSE Forum 2014, Thessaloniki, Greece, June 16-20, 2014, Selected Extended Papers*, volume 204 of *Lecture Notes in Business Information Processing*, pages 233–249. Springer, 2014.

[27] T. M. van Engers, R. van Gog, and K. Sayah. A case study on automated norm extraction. In T. Gordon, editor, *Legal Knowledge and Information Systems. Jurix 2004: The Seventeenth Annual Conference.*, Frontiers in Artificial Intelligence and Applications, pages 49–58, Amsterdam, 2004. IOS Press.

[28] R. R. Yeddula, P. Das, and S. Reddy. A model-driven approach to enterprise data migration. In J. Zdravkovic, M. Kirikova, and P. Johannesson, editors, *Conference on Advanced Information Systems Engineering (CAISE), Stockholm, Sweden*, volume 9097 of *Lecture Notes in Computer Science*, pages 230–243. Springer, 2015.

[29] N. Zeni, N. Kiyavitskaya, L. Mich, J. R. Cordy, and J. Mylopoulos. GaiusT: supporting the extraction of rights and obligations for regulatory compliance. *Requir. Eng.*, 20(1):1–22, 2015.