RC25346 (WAT1301-054) January 23, 2013
Computer Science

IBM Research Report

RTS - An Integrated Analytic Solution for Managing Regulation Changes and Their Impact on Business Compliance

Davide Pasetto, Hubertus Franke
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 208
Yorktown Heights, NY 10598
USA

Weihong Qian, Zhili Guo, Honglei Guo, Dongxu Duan, Yuan Ni, Yingxin Pan, Shenghua Bao, Feng Cao, Zhong Su
IBM Research Division
China Research Laboratory
Building 19, Zhouguancun Software Park
8 Dongbeiwang West Road, Haidian District
Beijing, 100193
P.R.China

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

RTS - An Integrated Analytic Solution for Managing Regulation Changes and their Impact on Business Compliance

Davide Pasetto Hubertus Franke
IBM T.J. Watson Research Center;
Yorktown Heights, NY 10598 USA
dpasett@us.ibm.com frankeh@us.ibm.com

Weihong Qian Zhili Guo Honglei Guo
Dongxu Duan Yuan Ni Yingxin Pan
Shenghua Bao Feng Cao Zhong Su
IBM China Research Lab
{qianwh, guozhili, guohl, duandx, niyuan, panyingx,
baoshhua, caofeng, suzhong}@cn.ibm.com

Abstract

Governance, Risk Management and Compliance are key success factors for corporations. Every company worldwide must ensure a proper compliance level with current and future laws and regulations, but managing the dynamic nature of the regulatory environment is a challenge, for both small and medium business as well as large corporations. Specifically the challenge is knowing and interpreting which regulations impact a particular business. Governments and standard bodies keep producing new, revised legislation, and businesses today rely on employees and consultants for tracking and understanding impact on their operations.

This paper introduces a novel prototype solution that addresses these concerns through the use of advanced text analytics. In particular the system is able to discover sources of regulatory content on the world wide web, track the changes to these regulations, extract metadata and semantic information and use these to provide a semantically guided comparison of regulation versions. Moreover, by leveraging the IBM DeepQA architecture, the solution is able to cross reference business objectives with the regulatory database and provide insights about the impact of new and revised laws on a company's business.

Categories and Subject Descriptors H.4.2 [*Types of Systems*]: Decision support

General Terms Design, Algorithms

Keywords text analytics, semantic, document processing, question answering

1. Introduction

In general compliance means conforming to a rule - for example a specification or standard. In the modern business world, "Regulatory Compliance" is the act of adhering to external laws and regulations, as well as internal corporate policies, procedures, and controls. This is the end goal of every corporation or public agency, that aspires to ensure that employees are aware of and take steps to comply with relevant laws and regulations. What is now known as corporate compliance is the result of many years of evolution and development. The laws covering businesses have grown over the years in size and scope just as the ways of dealing with these laws have grown more formal and complex. Regulation started slowly in the 19th century and picked up momentum in the ensuing years. Almost every regulation began as a response to individual scandals, and sought to address the underlying causes of each of these scandals. By the 1960s, with increasing complexity in both the business and regulatory arenas, the foundations of modern compliance began to emerge. This trend continued into the 1970s and 1980s, until it reached a tipping point with the release of the Sentencing Guidelines for Organizations[15] in 1991. Compliance programs existed well before these sentencing amendments, but the amendments gave these programs a major push into the mainstream of business.

Assessing whether a company's business practices conform to laws and regulations and follow standards and SLAs, i.e., compliance management, is a complex and costly task. Some software tools for compliance management do exist [10] under the umbrella of "Governance, Risk Management and Compliance" description; yet, they typically do not address the needs of who is actually in charge of assessing and understanding compliance, but rather focus on policy tracking and auditing to ensure a proper operational behavior. These systems help organizations understand their compliance level once the appropriate policies are defined and put in place, but do not help compliance officers to understand regulatory changes and their impact on the existing policies. *The cost of non compliance can be huge; for example the median fine [16] (legal cases between 2007 and 2010) for violating import-export regulations is $14,000,000, which lowers to $5,000,000 for environmental compliance, "only" $1,000,000 for each product safety or quality issue and rises above 40,000,000 dollars for competition and antitrust.*

This paper introduces RTS - Regulation Tracking Solution - an integrated analytic system that helps compliance officers discover which regulations they should comply to, track and understand changes and gain insight about their impact over the corporate compliance level. RTS is still a work in progress that utilizes bleeding edge technologies and capabilities both to reduce corporate risk levels, ensuring an on time understanding of laws and regulation changes impact, and unlock new business values by discovering what else is possible (compliant) within the current business environment.

The paper is organized as follows: Section 2 provides a high level overview of the solution, Section 3 details regulation change

tracking, followed by semantic comparison in Section 4. Section 5 describes the use IBM DeepQA pipeline to provide insight over the regulation database and finally Section 6 contains some concluding remarks.

2. Solution Overview

Inside a company, legal or compliance department regulation documents are usually processed by employees using a four step high level workflow described in Fig. 1. First, users are provided with documents for them to browse and explore. Second they determine certain topics or domains for a deep-dive. They can do search, read documents categorized by multiple angles (for example, by time, author or topic, etc.), and check trends. Third, users are able to compare target regulation documents and make sense of the comparative results for further action. In addition, for newly-released products, users need to ensure the compliance with all relevant regulation documents. To help the users fulfill these tasks, the team is designing a Regulation Tracking Solution which is architected into three layers (see Fig. 2):

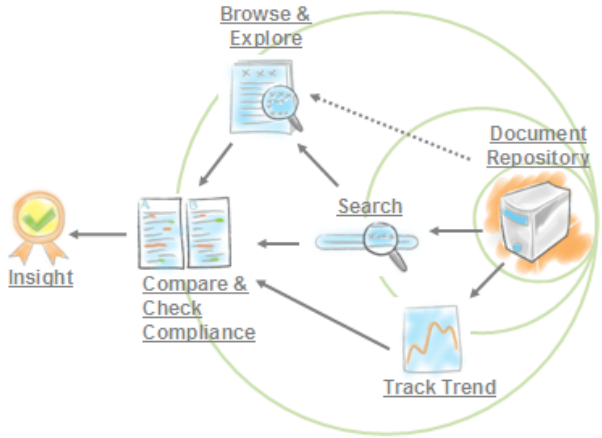


Figure 1. Laws and regulation document handling inside a company legal department.

Data layer Regulation documents are distributed by a variety of mechanisms such as database records or web sites. While it is an easy task to import structured data from existing databases, accessing publicly available regulation documents scattered on the web is still a challenge. It is obviously not realistic (and not even necessary) to crawl all the information on the web: there are mainly two kinds of regulation document sources that are valuable to end user: (1) documents that contain specific keywords and/or topics of the user's interests, (2) documents that are newly published in websites that a user is tracking. Based on these observations, it is possible to design a focused crawler able to leverage ranked results, returned by existing search engines, and re-filter them to select only specific documents. Once a website has been identified as a good source of regulatory content it can be crawled for further updates.

Analytic layer Raw documents collected by the data layer are usually in different format, like pdf, word, and webpage and the analytic layer provides an view across them. For example document normalization engines cleanse and segment the document content; a semantic engine extracts various metadata like authors, publishing time, and key topics; document categorization estimates the document similarity between each other and categorizes them into different clusters; the DeepQA [6] engine provides the decision support to the whole system.

Application layer Various kinds of applications can be developed based on the output of the analytic engines. This paper mainly focuses on four applications: (1) Trend analysis, which can help the users to easily consume large amount of regulation documents especially when entering a new domain. (2) Semantic search, which enables the users to find regulations most relevant for their business. (3) Semantic comparison, which can assist the user in comparing two or more regulation documents side by side at the semantic level: similar text with different meaning will be highlighted, and different text with identical meaning can be ignored. (4) Compliance checking, which is able to identify whether a new product violates existing regulations or whether a new regulation affects existing products, also providing the fine-grained evidences which explain why. All above applications are implemented and interacted in a visualized way.

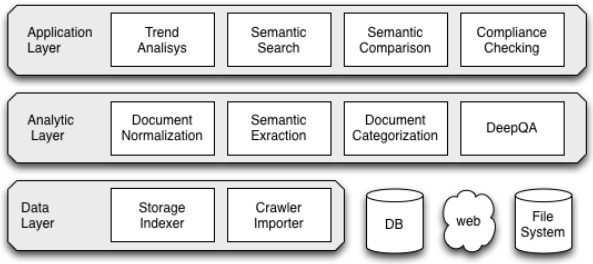


Figure 2. RTS system high level architecture: the solution is logically divided in three layers - data, analytic and application.

3. Regulation Tracking

Understanding changes in regulations is a complex and difficult task: governments and other regulatory bodies continuously modify and extend laws and regulations, changing definitions, applicability, exceptions and other details to better adapt to a constantly changing business environment. Each country, state and municipality has its own, slightly different, rules for every topic and activity, and understanding the differences between them is time consuming. Moreover, regulatory bodies often publish requests for proposals and draft new bills; companies that can react quickly enough, have the opportunity to analyse the impact on their own business and propose changes. The regulation tracking module provides a number of functions to deal with this information overload, allowing users to visualize interesting aspects of the regulatory data and search for applicable regulations. In this paper we detail two modules: trend analysis and semantic search.

3.1 Trend Analysis

One main task, that can help the users consume the large amount of regulation document more easily, is analyzing the trend of regulation data corpus. The goal is to understand the content of regulation documents and how the content evolves and changes over a specific dimension, such as time or space. This is extremely useful both for users who enter a new domain and for people that look ahead, trying to understand where regulators are going and prepare in advance.

This module uses an interactive visual text analysis tool, called TIARA [11] (Text Insight via Automated, Responsive Analysis), to visually summarize results of regulation trend analysis. Figure 4 shows a time-based, domain-oriented visual text summary of four regulation domains. Each colored layer represents a regulation domain. Each layer is depicted by a set of keyword clouds, summarizing the content of regulation documents and the content evolution

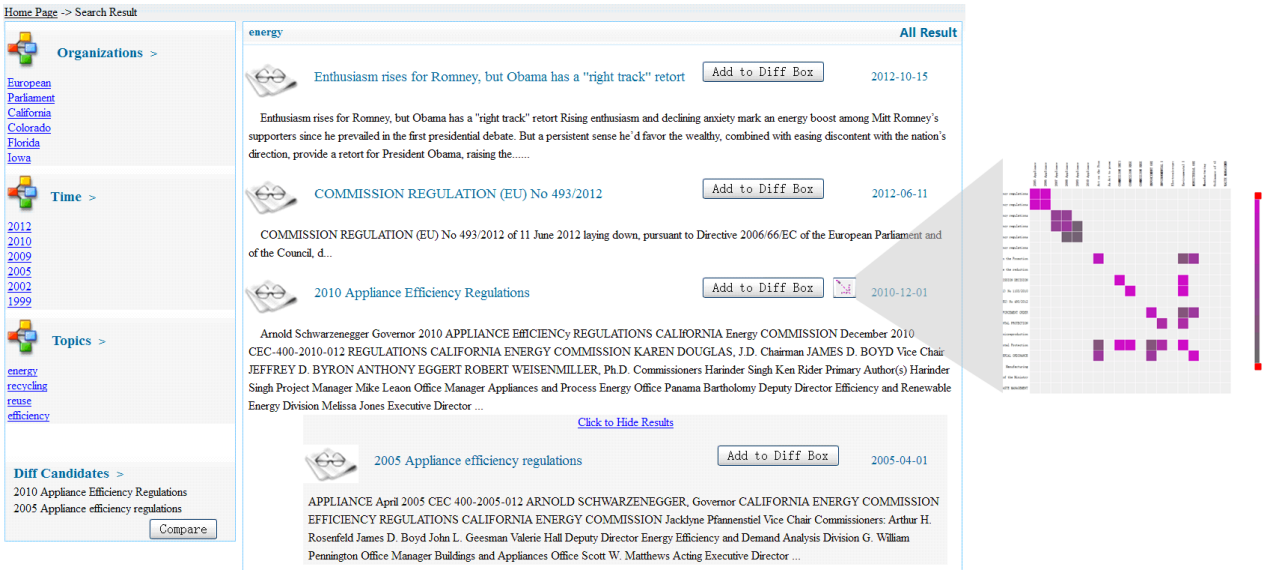


Figure 3. Document semantic search visualization where results are grouped with related documents. On the left the navigation filters operate on document metadata, while on the right regulation correlation index is displayed for each group of related document.

over time. These keyword clouds are automatically extracted by using a Latent Dirichlet Allocation [3] (LDA) model. The width of a layer at a time point encodes the number of regulation documents covering the domain at that time. This highly visual summary of regulation trends, together with keyword clouds laid out inside the layers, provides the users a good overview of what's going on in the space.

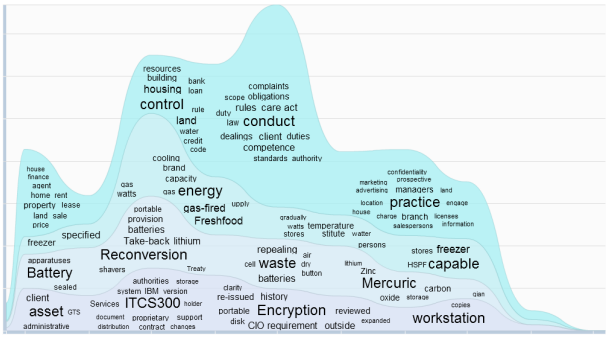


Figure 4. Trend analysis and visualization for regulation documents. High level topics are layered and their key content is shown through word clouds. The graph shows the evolution of the regulated topic over time by counting the number of articles and analyzing how the used concept taxonomy changes.

Starting from the high level overview of regulation trends, the user can zoom on specific areas for more details. Often only a subset of regulation keywords can be displayed within a regulation domain due to the limited screen size, so users may request more details about a specific domain to help make decisions, TIARA allows users to interactively zoom in/zoom out on a selected domain using a fish-eye view technique. While looking at the trend of regulation documents, users often need to know the exact number of documents at a certain time among different domains. TIARA uses the needle technique to report the numbers of regulation documents for a selected data-point. After finding interesting keywords or topics, users can trigger a semantic search simply by clicking a keyword in a domain.

3.2 Semantic Search

The search page is accessible both from the topic trend visualization and through a standard expression based query string; it contains two panels: navigation and output regulation list (see Figure 3). Inside the navigation panel a number of features (such as organization, time or topic) extracted from the document that match the search criteria are exposed to enable further refining. The resulting regulation documents list is divided into different groups, each satisfying a distinct semantic meaning; each group is represented by the most up to date document.

Users can look for the topic of interest efficiently by going through the titles or summaries of these documents. The user has also the ability to examine the content of particular document group. For example in Figure 3 the regulation "2010 Appliance Efficiency" is grouped with its older version "2005 Appliance Efficiency". A number of regulations belonging to a group can be added to the "diff. candidates" list to access their semantic comparison, detailed in Section 4.

In order to help understand related regulation documents more intuitively, an adjacency matrix representation is provided (see right part of Figure 3). In the adjacency matrix, documents are laid out as rows and columns and document similarity measure are represented in the corresponding cell in the matrix: different colors encode different similarity score. Each colored cell highlights two comparable documents and how similar they are. Matrices have two advantages which make them more readable than basic document list: (1) as documents are represented both in rows and columns, the relationship between any two documents is very clear and (2) rows and columns of matrix can be reordered (manually or automatically) to improve readability and visually highlight clusters of document regulating the same topic.

4. Semantic Comparison

Governments and regulatory bodies keep updating their regulatory content to cope with constantly changing requirements of business practices. Each change, no matter how small, can impact an organization's compliance level and must be studied in detail. Unfortunately, existing tools cannot effectively distinguish between purely

cosmetic changes or semantic content changes, nor can they highlight whether a specific difference has business importance or can potentially be ignored. The goal of the semantic comparison function is exactly this: help users compare regulation documents at a semantic level, not at the usual text level, and highlighting the sections they should study in detail.

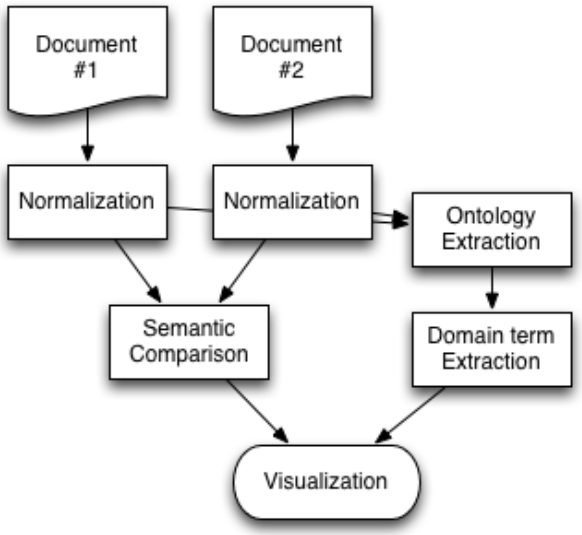


Figure 5. Semantic Comparison Overview.

Performing semantic comparison of regulation documents is a complex task that requires a number of steps: converting each input document into a structured representation (often called "document normalization"), semantically aligning documents and applying Myers [12] diff algorithms on paragraph, sentence, and at word level producing an XML encoded comparison document. Finally, the resulting structured document must be visualized in an intuitive way. The overall process is depicted in Figure 5

4.1 Document processing

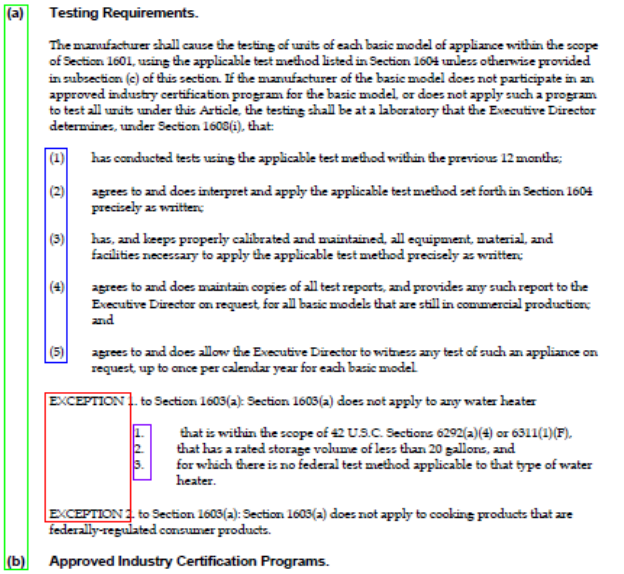
Handling generic document normalization requires a number of steps, such as converting each input document into a structured XML representation; extracting texts and formatting markups; segmenting each document into chapters, sections, paragraphs and sentences; tagging each part with its content; running morphologic analysis; and extracting ontology phrases and domain terms. Multiple of domain concepts and terms are employed inside the regulatory documents. These concepts and terms are very important semantic indicators for deep semantic difference analysis in regulation comparison. In order to provide better semantic comparison and effective navigation in RTS, the tool extracts also domain-specific ontology tree and topic terms from the documents.

4.1.1 Format Conversion

Most regulation documents are published using Adobe PDF, Microsoft Word, or HTML file formats. To normalize their contents we can use either TIKA [2], an Apache open source tool, or Nuance OmniPage [13] to convert various file formats to a standard XML representation. TIKA based conversion recognizes only page breaks and line breaks while NUANCE based conversion can output a large number of structural and presentation information, such as font and color.

4.1.2 Document segmentation

To be able to perform semantic comparison, it is necessary to obtain the real document structure. Most important structural features are chapters, sections, paragraphs and sentences.; furthermore each part should be labelled as "definition", "scope", "exception", "method", or "limitation", etc., concepts that are common in regulation documents, and are necessary to help understanding the regulation contents. Unfortunately these document are unstructured and meant to be consumed by humans, not machines. Markup tags in the XML output that TIKA and NUANCE produce are mainly indicators of page and lines, which do not directly map to boundaries of semantic units like chapters, sections, subsections, paragraphs and sentence. To identify these semantic boundaries, the solution applies several heuristics:


(a) **Testing Requirements.**

The manufacturer shall cause the testing of units of each basic model of appliance within the scope of Section 1601, using the applicable test method listed in Section 1604 unless otherwise provided in subsection (c) of this section. If the manufacturer of the basic model does not participate in an approved industry certification program for the basic model, or does not apply such a program to test all units under this Article, the testing shall be at a laboratory that the Executive Director determines, under Section 1608(i), that:

(1) has conducted tests using the applicable test method within the previous 12 months;

(2) agrees to and does interpret and apply the applicable test method set forth in Section 1604 precisely as written;

(3) has, and keeps properly calibrated and maintained, all equipment, material, and facilities necessary to apply the applicable test method precisely as written;

(4) agrees to and does maintain copies of all test reports, and provides any such report to the Executive Director on request, for all basic models that are still in commercial production; and

(5) agrees to and does allow the Executive Director to witness any test of such an appliance on request, up to once per calendar year for each basic model.

EXCEPTION 1. to Section 1603(a): Section 1603(a) does not apply to any water heater

1. that is within the scope of 42 U.S.C. Sections 6292(a)(4) or 6311(1)(F),
2. that has a rated storage volume of less than 20 gallons, and
3. for which there is no federal test method applicable to that type of water heater.

EXCEPTION 2. to Section 1603(a): Section 1603(a) does not apply to cooking products that are federally-regulated consumer products.

(b) **Approved Industry Certification Programs.**

Figure 6. Document structure clues example.

1. Long documents usually contain a "table of contents", with starting page numbers or HTML anchors. Using these clues it is possible to split the whole document into chapters and sections.
2. Regulation documents are usually composed by many sections (called articles). Numbered lists and bullet lists follow specific style instruction, such as in the example shown in Figure 6. Numbered lists like (a), (b) form the top-level structure of this chapter. Then digits enclosed in parentheses form the second-level lists, etc. The system scans the whole document to learn the nesting styles, then segments the document into finer-grained subsections.
3. Within subsections it is possible to leverage line breaks, capitalization of lines, and some special clue words to identify paragraph boundaries. Other techniques include taking consecutive line breaks, lines whose lengths are below average length, and lines starting with clue words like "Phone:" or "Fax:", as paragraph boundaries.
4. After identifying paragraphs, it is possible to segment each paragraph into sentences by using punctuation marks, word capitalization, and abbreviation list.

4.1.3 Linguistic-based Ontology Extraction

Key domain concepts and terms are often defined inside the regulation document itself, to avoid any misunderstanding while reading

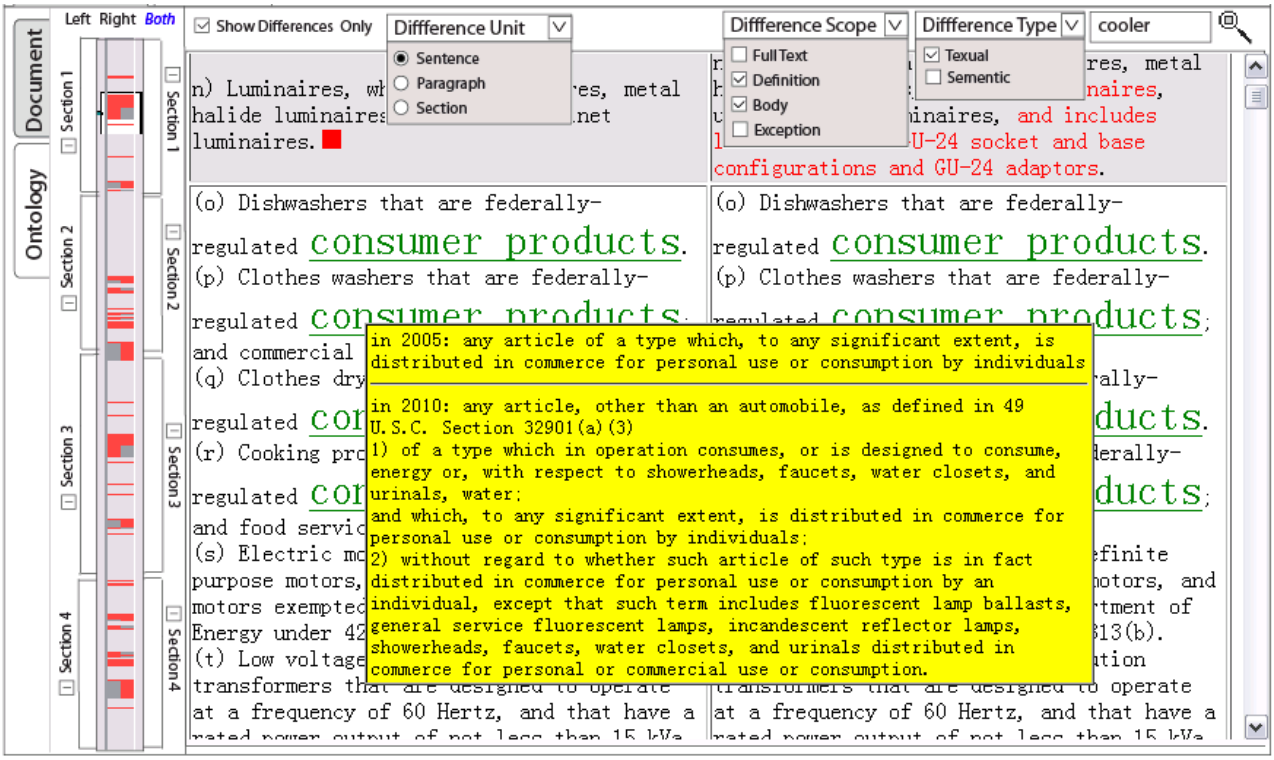


Figure 7. Document semantic comparison visualization.

the text. For example, inside the California Appliance Efficiency regulation it is possible to see that *"Blast chiller" means a refrigerator designed to cool food products from* 140° *F to* 40° *F within four hours. "Buffet table" means a commercial refrigerator, such as* ..... All the concepts are defined using some document specific templates, such as "A means B that ...", "A means B which ..." or "A means B for ..." and so on.

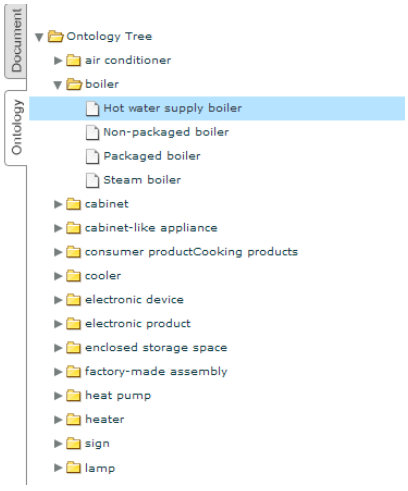


Figure 8. Domain ontology tree for California appliance efficiency regulation.

The tool employs linguistic-based pattern analytics methods to extract domain ontology from the regulation documents. The first step is to detect these definitions from within the documents. Then it is possible to extract the concept terms from these definitions using only linguistic rules. Finally, the solution identifies the semantic relations among these concepts by using pattern-based semantic analysis.

For example, given the definition *"Bottle-type water dispenser" means a water dispenser that uses a bottle or reservoir as the source of potable water*, the system first extracts the two concepts "bottle-type water dispenser" and "water dispenser" from it. Then it identifies that "bottle-type water dispenser" is a specific incarnation of the super category "water dispenser". Figure 8 shows the domain ontology tree extracted from the California appliance efficiency regulation documents, consisting of 43 nodes.

4.1.4 Domain term categorization with latent semantic association

Domain terms provide rich semantic hints for deep text analysis; however it is challenging to recognize such terms from the available document structural information. To overcome this challenge the system uses a domain term extraction and categorization method with latent semantic association. First, documents are tokenized and each word is tagged with the part-of-speech [5] by using the popular natural language processing toolkit OpenNLP [1]. Then all the noun phrases (for example with frequency ≥ 3) are detected from the documents using linguistic rules. These noun phrases are further ranked according to their frequency in the dataset. Finally, the top n high-frequent key terms from the ranking list are selected.

In order to capture the semantic association among domain terms, they are further categorized into semantic groups using LaSA [7]: latent semantic association model. A LaSA model can be considered as a general probabilistic topic model that can be learned from the unlabeled corpus using popular hidden topic models such as LDA [3] (Latent Dirichlet Allocation) or pLSI [8] (probabilistic Latent Semantic Indexing).

Table 1. Domain term samples in battery recycling domain

Aspects	Sample words
Recycle	recycling process, recycling, battery recycling process, minimum recycling efficiencies, recycling steps, recycling efficiencies, recyclers, disposal, recycling facility, collection
Battery	battery, waste batteries, battery packs, storage batteries, accumulators, nickel-cadmium batteries, lead-acid batteries, waste battery types
Battery-Content	content, lead content, water content, cadmium content, cadmium content, average Pb content, chemical composition, compound, element

The LaSA-based domain term categorization constructs a virtual context document for each domain term to describe its latent semantic distribution. The virtual context document is composed of all the available internal lexical clues and the external context clues, such as the component words, headwords and the co-occurrence adjacent noun words/terms in the documents. Then it computes a topic model for these virtual context documents. In order to effectively categorize domain terms, it employs the popular hidden topic model LDA to learn the semantic association among the domain terms from their virtual context documents. In the topic model building, σ is a topic model with k topics, $\sigma = \sigma_1, \sigma_2, ..., \sigma_k$.

σ learns the posterior distribution to decompose domain terms and their virtual context documents into topics. Given a domain term x_i and its virtual context document vdx_i , each vdx_i is grouped into one topic σ_j by choosing the topic model with the largest probability of generating vdx_i. The semantic categories of domain terms are inferred with the generated topic model; Table 1 shows an example domain term category output for documents in the battery recycling domain.

4.2 Document Comparison

To compare two regulation documents, the system first aligns chapters, sections, subsections, and paragraphs. This is extremely useful for regulation documents, which are usually long and can contain many parts. Then a mature tf-idf [14] statistic is used to reflect how important each word is with respect to each semantic part. This statistic is boosted for ontology phrases and domain terms, since these have been previously discovered as important information inside the document. These combined weighting factors are summed and stored for each text portion where the two documents differ.

4.2.1 Semantic unit alignment

During document preprocessing the system extracted chapter titles and article titles. These titles are used as the basis for alignment. Aligned sections are put into a comparison queue. If a section can't be aligned it is simply stored as either inserted or deleted parts in the final output.

4.2.2 Chunk difference ranking

The solution uses an enhanced Myers diff algorithm with two-layer support to generate textual diff chunks for the two documents. In the top layer sentences are used as comparison units, thus identical sentences are aligned, and annotated with three chunk difference type: insertion, deletion, and modification. Myers diff is extended with a fourth difference type: position switch. Once the basic diff is completed, the weights of all different words or phrases is summed and stored as the difference score for the chunk.

4.2.3 Difference semantic tagging

Chunks differences are also tagged with their semantic type by leveraging the paragraph tags added during document preprocessing: all semantic tags (such as "definition", "scope", "exception", or "limitation", etc.) whose scope overlaps the changes text are added as tags for the specific section. This allows the system to support extensive semantic difference highlighting. For example, if one domain term definition changed, by tagging it across the entire text we can highlight every use as a potential difference even if the text referring to the term did not change.

4.3 Semantic Comparison Visualization

To consume the document comparison result, the team designed an interactive diff visualization for regulation documents (see Figure 7). The main view displays comparison result of two documents, side by side and structurally aligned by the various document subdivision units: section, paragraph and sentence. Individual words are color coded to highlight changes: red represent different content, black similar content and gray missing text portions.

The tool can display three types of differences:

- basic typographical changes, like any standard diff tool.
- typographical changes with semantic information, which highlights also high level concepts that are different between the two document even when there's no text change.
- semantic differences, which shows only the semantically significant changes.

To help the users better explore the documents, the interface provides several navigation options, including the ontology tree extracted from the document and keyword search capability.

5. Compliance Checking

Laws and regulations play an important role in ensuring that companies follow appropriate practices and procedures while running their business. The cost of violating regulations can be extremely high; regulations are voluminous, and verification and cross referencing is becoming increasingly complex. It is quite difficult and very time consuming for human beings to, given a specific product or requirement, manually locate all related provisions and obligations, from various laws and regulations, that could be applicable to the situation. Several researched approached the problem by proposing methods to formalize provisions inside text-based regulations in order to simplify regulation compliance checking [4, 9]. While this has been done in some highly specific environment, most topics are regulated through the text based description of situation and regulations. Unfortunately, automatically understanding requirements that are represented in the natural language, and build a semi formal set of provisions representing them, is challenging and, when possible, also computational intensive. For example suppose a company wants to start sell a product, a battery with a zinc anode and alkaline electrolyte manufactured in Ohio in 2012, in Wisconsin. The compliance officer task first task is finding all the provisions that could be applied to this situation. By using basic keywords like "zinc anode", "alkaline electrolyte" and "Wisconsin", the regulators could find relevant sections from the Wisconsin legislative documents. However, the attorney need to read througly all the related sections, as well as some section not identified by the search, to find the rule that applies in this specific situation: 100.27.(3) "Zinc carbon batteries. No person may sell or offer for

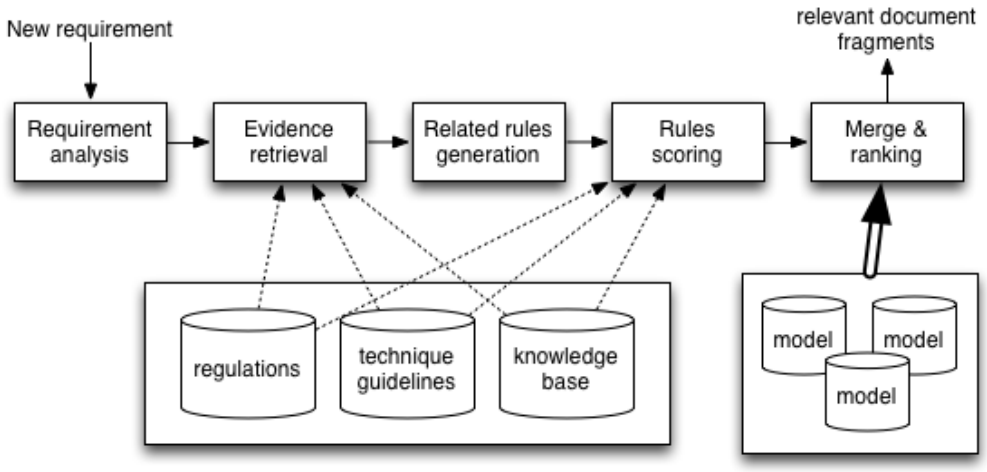


Figure 9. DeepQA architecture for regulation insight.

sale a zinc carbon battery that is manufactured after July 1, 1994, unless the manufacturer has certified to the department that the zinc carbon battery contains no mercury that was intentionally introduced."

The RTS system reuses IBM's DeepQA framework [6] to automatically locate the set of provisions that could be applied to a specific requirement, understand the connection between the provisions, extract and rank supporting evidence and present the user with a concise answer. The DeepQA framework was introduced in the supercomputer codename "Watson", who competed in the popular Jeopardy! game and won against the all times best players. Figure 9 shows the DeepQA system architecture when applied to laws and regulation processing; it consists of five components:

- Requirement Analysis. Given a new requirement, formulated as a natural language question, this component performs semantic analysis text to extracts the facts that should be demonstrated. The analysis will identify the key concepts and key relationships in the requirements, which are the foundation for understanding the real requirement. For example, with respect to the above example, the module will identify the Wisconsin as a location, 2012 as the production year, battery as a product along with its characteristic attributes such as zinc anode and alkaline electrolyte.
- Evidence Retrieval. Starting from the understanding of the requirement, the knowledge-base (which contains a background knowledge plus the regulatory documents located through the search and tracking mechanisms already included) is mined looking for related evidence sources such as portions of bills, sentences inside guidelines, and etc. This module returns a set of potentially important passages and it is similar to a keyword search that as to guarantee an high recall: all sections that contain information potentially useful for building the output should be returned.
- Related Rule Generation. The text portion from the previous step contain a set of provisions, but not all of them will apply to the specific requirement. For example, in Wisconsin legislative document, section 100.27 overall contains 18 sub provisions, but rule 100.27.5 is not important for the original query. This component will segment each rule from the related sections, and insert each segment as candidates for further analysis.
- Rule Scoring. This component embeds a set of scorers that use the retrieved evidences to determine the probability (in various dimensions) of whether the candidate provisions can be applicable to the requirement.
- Merge & Ranking. The module merges all scores to generate a final confidence value by applying a few machine learning models, and then ranks the provisions according to their final confidences. The provisions that have a large enough confidence level are the final answer, along with the supporting evidence (text passages) that define them.

6. Conclusion

This paper describes the software architecture of the RTS prototype, a complete turn-key solution for discovering, managing, analyzing and providing insight over laws and regulations. The motivation for this effort comes from the fact that compliance is increasingly important for modern companies: on one hand government and other regulatory bodies constantly modify regulations to cope with new challenges arising in the business environment. On the other hand the risk associated with non compliance is so high that business performance, especially for publicly traded companies, is assessed also in terms of how well a corporation manages its overall compliance and the associated risks.

The RTS solution aims at providing a semantically rich environment, leveraging the large number of innovative techniques for natural language processing and automatic reasoning that were devised in recent years, and applying them to the problem of compliance complementing the subject matter experts and company attorneys improving their efficiency.

RTS is still work in progress and it is now used internally inside IBM for managing environmental and chemical compliance issues. Future work includes:

- a throughout evaluation of the benefits and drawback of automation in legal document tracking and analysis
- an analysis of which other technologies and techniques for natural language processing might be applicable to this problem space
- a study about how multiple languages impact regulatory compliance tasks, including how and when to use automatic translation and if and how taxonomies and ontology align in different languages.

References

[1] Apache Fundation. OpenNLP. http://opennlp.apache.org/, .

[2] Apache Fundation. TiKA. http://tika.apache.org/, .

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=944919.944937.

[4] K. R. Bouzidi, B. Fies, C. Faron-Zucker, A. Zarli, and N. L. Thanh. Semantic web approach to ease regulation compliance checking in construction industry. *Future Internet*, 4(3):830–851, 2012. ISSN 1999-5903. doi: 10.3390/fi4030830. URL http://www.mdpi.com/1999-5903/4/3/830.

[5] E. Brill. Some advances in transformation-based part of speech tagging. In *Proceedings of the twelfth national conference on Artificial intelligence (vol. 1)*, AAAI '94, pages 722–727, Menlo Park, CA, USA, 1994. American Association for Artificial Intelligence. ISBN 0-262-61102-3. URL http://dl.acm.org/citation.cfm?id=199288.199378.

[6] D. A. Ferrucci. Introduction to "this is watson". *IBM Journal of Research and Development*, 56(3):1, 2012.

[7] H. Guo, H. Zhu, Z. Guo, X. Zhang, and Z. Su. Product feature categorization with multilevel latent semantic association. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1087–1096, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. doi: 10.1145/1645953.1646091. URL http://doi.acm.org/10.1145/1645953.1646091.

[8] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM. ISBN 1-58113-096-1. doi: 10.1145/312624.312649. URL http://doi.acm.org/10.1145/312624.312649.

[9] 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 *Proceedings of the 27th International Conference on Conceptual Modeling*, ER '08, pages 154–168, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-87876-6.

[10] G. T. Lau, S. Kerrigan, E. Engineering, G. Wiederhold, E. Engineering, and K. H. Law. An e-government information architecture for regulation analysis and compliance assistance. In *ICEC'04: Sixth International Conference on Electronic Commerce*, pages 461–470, 2004.

[11] S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian. Interactive, topic-based visual text summarization and analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 543–552, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. doi: 10.1145/1645953.1646023. URL http://doi.acm.org/10.1145/1645953.1646023.

[12] E. W. Myers. An o(nd) difference algorithm and its variations. *Algorithmica*, 1:251–266, 1986.

[13] Nuance. OmnuiPage. http://www.nuance.com/omnipage/.

[14] K. Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 1972. doi: http://dx.doi.org/10.1108. URL http://www.soi.city.ac.uk/ ser/idf.html.

[15] U.S. Sentencing Commission. 2009 Federal Sentencing Guidelines Manual. http://www.ussc.gov/2009guid/tabcon09.htm, 2009.

[16] U.S. Sentencing Commission. Understanding Compliance Risk in Emerging Markets. http://www.exbd.com/, 2011.