# Compliance Activator

## 1 Motivation

Organizations are increasingly confronted with all sorts of regulations to comply with in order to avoid penalties, image problems or the like. For responsible people it is getting more and more difficult and time consuming to identify regulations and monitor relevant changes. IT-based data analytics procedures are promising candidates to support compliance managers in their work. Compliance Activator provides such a service, based on innovative semantic analysis technology [Schmidt et al. 2014].

## 2 Architecture

The solution consists of the data source integration layer, the analysis layer and the action layer (fig. 1). The first two layers form the system core, mainly represented by the Java-based the standard software iQser GIN Server [iQser 2015]. The underlying big data technology stack includes Hadoop, Cassandra and ElasticSearch, allowing for both horizontal and vertical scalability.

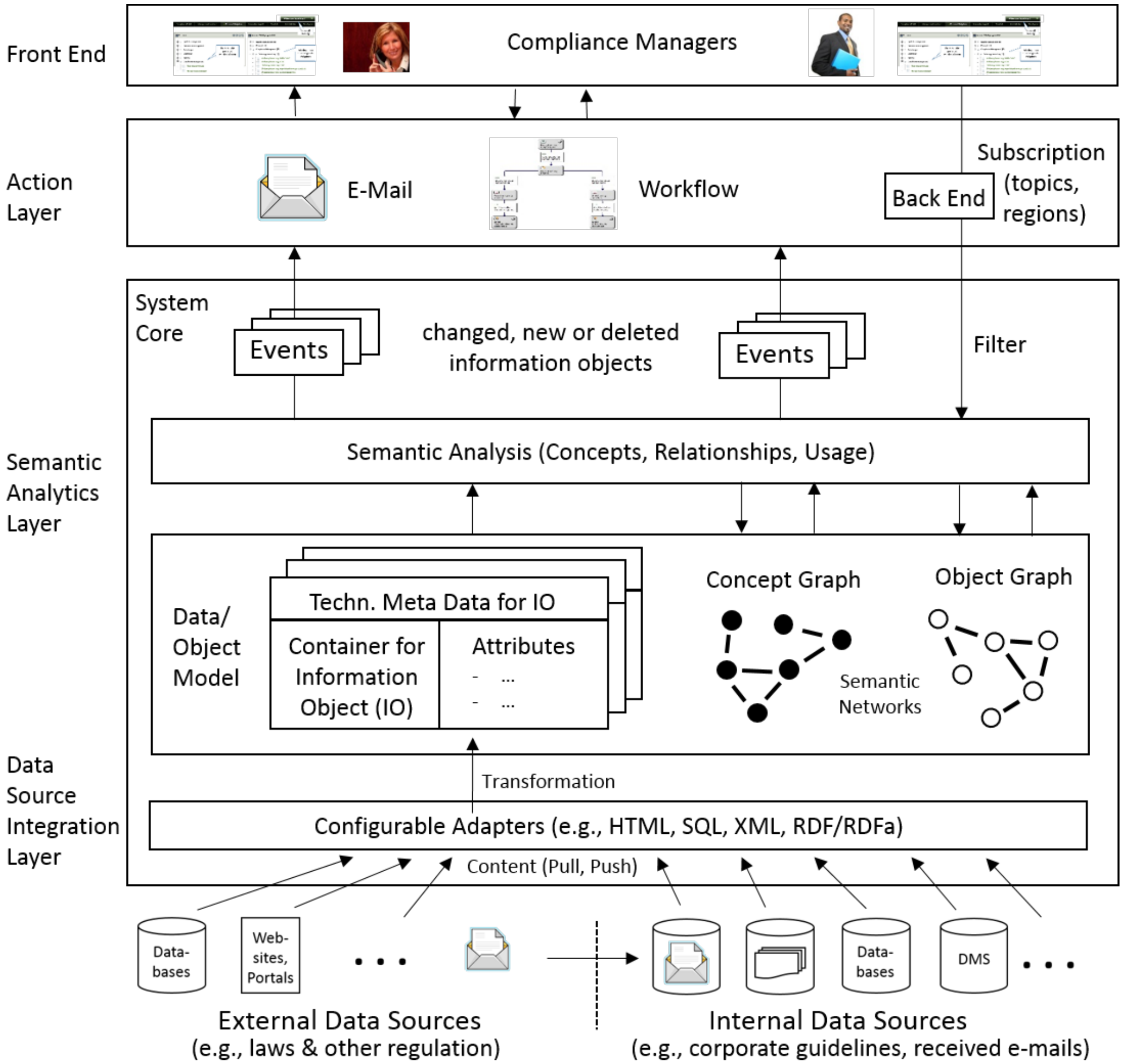
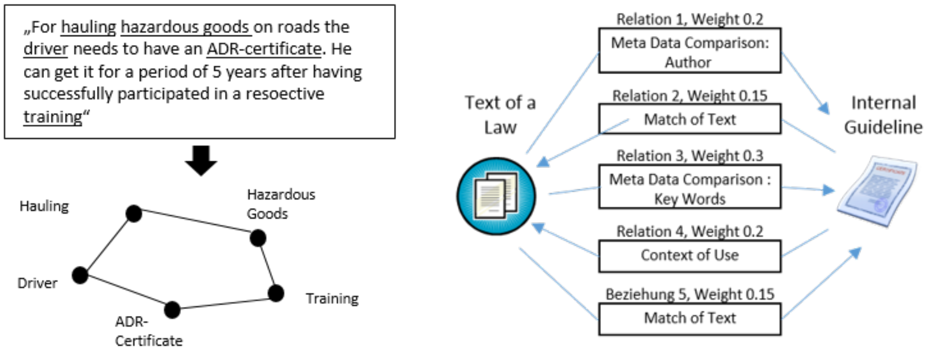


Fig. 1: Architecture of Compliance Activator

The **Data Source Integration Layer** connects sources of structured and non-structured information with heterogenous formats, using adapters for example for query and description languages, messaging and social network services and all sorts of files supporting Standard Content Management Interoperability Services (CMIS). Currently configurable adapters are available as Open-Source plugins for HTML, Structured Query Language (SQL), Extensible Markup Language (XML), RSS, Resource Description Framework in Attributes (RDF/RDFa), MS Exchange, IMAP, Facebook or LinkedIn (a list is available at Sourceforge [Sourceforge 2015]). Content from the various sources is transformed into a unique logical model [Smolnik & Wurzer 2009]. The transformation includes storing content as information objects (IO) in containers, together with technical meta data like Uniform Resource Identifier (URI), origin, modification date and attributes like type (text, photo etc.). Thus container objects already include semantic descriptions of their data. From the concrete entities the system is able to derive classes and build an abstract data model. The subsequent semantic analysis (see next paragraph) develops relationships between information objects. This procedure allows integrating arbitrary structured and unstructured external data sources (public websites, social networks, chargeable databases etc.) as well as internal data (e.g., e-mail messages, database records, documents on fileservers) into the analysis and relating information objects semantically.

The **Semantic Analytics Layer** applies a combination of methods for building and maintaining semantic networks [Smolnik & Wurzer 2008, Wurzer 2008]. The **concept analysis** uses natural language programming methods in order to extract concepts and their relationships from data (for example see [Basili et al. 2012]). Significance measures the importance of a concept. It indicates how characteristic a concept is for the topic of the text. Auxiliary verbs and adjectives for example are of low significance in contrast to names and nouns. Co-occurence measures collective existence of highly significant concepts within sentences. Concept analysis results in a concept graph (see fig. 2, left). This graph is enriched and maintained by the following steps.



„For hauling hazardous goods on roads the driver needs to have an ADR-certificate. He can get it for a period of 5 years after having successfully participated in a resoective training“

Fig. 2: Concept Graph (left) and Content Graph (right)

**Syntax and pattern analysis** identifies relationships between information objects and is pictured in the content graph (see fig. 2, right). Syntax analysis connects objects by (key) attributes, for example the sender of an e-mail message or author of a word document or a blog post. This allows bundling all information created by a person being expert in a certain domain like compliance regulation for international trade. Pattern analysis uses text mining methods to reveal coherence of content objects based on topics in their free texts. It takes concepts of the concept graph for queries to retrieve similar information objects. Similarity values between 0 (totally different) and 1 (intentionally connected by user) serve to evaluate thematic relatedness. **Analysis of usage** refers to the fact that – like with words in language – the use of information objects changes their significance over time. Main drivers of this change are their relevance for the user and the context in which they are used. Monitoring user interaction on information objects like creating, modifying, deleting allows recalculating relevance of and relationships between information objects. Thus connections can increase, decrease or even diminish over time. If, for example, a logistics manager regularly accesses a list of hazardous goods, then updates an entry in a database for material classification and finally sends out an e-mail with a respective subject the system reasons that the involved information objects are related to each other.

The **Action Layer** provides functions triggered by the events thrown by the analytics layer. The concrete design is a matter of requirements. Functions can include sending e-mails to inform people about a change in a content they might be concerned with, e.g., compliance managers for particular regulation areas and regions. Advanced functionality can comprise more sophisticated workflows. The system can also tell the user more precisely what part of a content has changed, also provide him with the former version and present internal information which needs to be checked for the necessity of being updated. Another feature is to run internal information against a best practice document (master) in order to find differences and missing aspects. An IT security manager can for example upload a best practice IT security policy to the system. Semantic Analytics Layer analyses it and retrieves all sorts of information chunks related to its content in the connected internal systems. Based on the results of the system's comparison with the master document the user can identify for example gaps in the company's IT security guidelines and decide what to do.

**Initial input and update** is accomplished as follows: After connecting the data sources the systems starts transforming and organizing the content and over time dynamically enriching and updating the derived data model with the mechanisms described above. Defining data models, ontologies or semantic annotations in advance is not necessary [Smolnik & Wurzer 2009, S. 4]. Rather the system successively builds an ontology for the user on the content objects, derived from the concrete entities, e.g. an ontology on data protection. In contrast to competing solutions this means that results of the analysis are not depending on the knowledge of the data modeler, but emerge from the analysed data itself. The system recognizes changes in the tethered sources through time-controlled crawling or ad hoc queries by the user (Pull). It is also possible for sources to actively notify modifications via service APIs of databases or calling mechanisms of web content providers (Push). In both cases changes undergo the semantic analysis modifying the semantic networks for example by new or deleted content objects and new or recalculated relationships. Each of these modifications causes an event, which is characterized by attributes like type (new content, content change etc.) and the affected information object, and triggers subsequent activities.

The **solution supports many languages**, not only with respect to the user interface, but also regarding the content to be analysed. This is facilitated by a machine translation option both for customer content and master documents to run the analysis against. Customers can use the service for example to automatically translate their office documents from German to English and then let the semantic analysis find relevant content matches in external data sources in English. Machine translation is currently available for German, English, French, Russian, Spanish, Portuguese and Chinese and vice versa.

### 3 Potential for multiple Use

Data Source Integration and Semantic Analytics Layer form the core of the solution. As content adapters and analytics are independent from industries, domains etc. the technology promises to be beneficial for many use cases beyond compliance management. For example lawyers, when preparing for a lawsuit, can easily get information relevant for their case from all connected sources like common law databases, public websites, social network and the like. Same holds for financial analysts who want to put together exposés about enterprises for investors. To generate a new application for a certain domain (e.g., financial analysis), it is only necessary to connect the respective data sources of the new topic (e.g., enterprise databases like MarketLine Advantage, stock exchange websites and financial magazines) to the semantic middleware and enable it as a new subscription in the backend of the application. The new content is being processed and permanently updated as described above, with analysis results being delivered to the customer (e.g., financial analyst) as subscribed by him.

Furthermore it is easy to create content adapters for formats not supported so far. It is also possible to enhance the analytics engine by adding specialized analysis stages. This makes the core pretty interesting for many real world problem scenarios related to big data and data consolidation and harmonization (also see next section).

## 4 Product Launch

The Compliance Activator will be launched by Analytical Semantics AG, a corporation currently in the course of formation. The website will go public under www.analytical-semantics.de after the corporation has been established. The application is available for iOS, Android and HTML 5. It is currently being implemented in globally active large enterprises (e.g., automotive) and in small and medium-sized enterprises (e.g., regional power suppliers).
There are also projects running for implementing a middleware called Enterprise Content Service Bus (ECSB), based on the described technology. The ECSB analyses all internal data sources and suggests a semantic scheme of the content and a mapping with business objects in Master Data Management (MDM) and Transaction Data Management. The user adapts the schemes according to his business (process) needs. The system then generates a first version of an overall MDM scheme. It permanently monitors structures and content of the source and gives alerts in case of changes. The MDM layer stores data and dynamically provides it for applications. It also comprises the possibility to create new content objects and to manipulate structure, fields and data of existing content objects. Those changes lead to a new analysis of the sources resulting in adapted semantic schemes, again subject to clearing and approval by the user. System and data owners thus are informed of changes in the sources on the ECSB layer and there can easily approve or reject them. This cyclic procedure assures that applications using those data objects can instantly react on changes in the data supplying systems and stay operational without disruption. Hence the ECSB allows a smooth lifecycle management and migration of heterogenous data sources into a more consolidated data architecture with parallel, yet consistent operation of legacy systems, systems to be eliminated and new systems.

## References

[Basili et al. 2012] *Basili, R., Stellato, A., Previtali, D., Salvatore, P., Wurzer, J.,* Innovation-related Enterprise Semantic Search: the INSEARCH experience, In: Proceedings of the 6th IEEE International Conference of Semantic Computing, Palermo 2012, 194-201.

[iQser 2015] *www.iqser.de*, letzter Zugriff am 31.07.2015.

[Schmidt et al. 2014] Schmidt, W., Braun, F., Kramm, A., Wurzer, J. Semantik-basiertes Compliance-Portal, in: HMD Praxis der Wirtschaftsinformatik 51 (2014) 3, Springer, Heidelberg, S. 293-306, DOI 10.1365/s40702-014-0031-2.

[Smolnik & Wurzer 2008] *Smolnik, S., Wurzer, J.,* Towards an automatic semantic integration of information. In: Proceedings of the International Conference on Topic Map Research and Applications (TMRA 2008), Leipzig, 2008, 169-179.

[Smolnik & Wurzer 2009] *Smolnik, S., Wurzer, J.*, Wissen dynamisch organisieren. Einsatzfelder einer automatischen semantischen Analyse verteilter Informationen. In: Konferenzband zum 11. Kongress KnowTech - Konferenz für Wissensmanagement, Bad Homburg, 2009.

[Sourceforge 2015] *www.sourceforge.net*, *Search item: iqser*, last access 2015-08-01.

[Wurzer 2008] *Wurzer, J.*, New approach for semantic web by automatic semantics. In: Accepted papers for the 2nd European Semantic Technology Conference 2008, Wien.