Defining and Retrieving Themes in Nuclear Regulations

Nicolas Sannier* ** and Benoit Baudry**
* EDF R&D – STEP, 6 Quai Watier BP49
78401 Chatou cedex, France
{nicolas.sannier, n.thuy}@edf.fr
** Inria, Campus Universitaire de Beaulieu,
35042, Rennes Cedex, France
{nicolas.sannier, benoit.baudry}@inria.fr

***Abstract*— *Safety systems in nuclear industry must conform to an increasing set of regulatory requirements. These requirements are scattered throughout multiple documents expressing different levels of requirements or different kinds of requirements. Consequently, when licensees want to extract the set of regulations related to a specific concern, they lack explicit traces between all regulation documents and mostly get lost while attempting to compare two different regulatory corpora.*

This paper presents the regulatory landscape in the context of digital Instrumentation and Command systems in nuclear power plants. To cope with this complexity, we define and discuss challenges toward an approach based on information retrieval techniques to first narrow the regulatory research space into themes and then assist the recovery of these traceability links.

Keywords: Regulatory requirements, theme organization, requirements traceability, information retrieval, domain practice

I. INTRODUCTION

Software systems designed to perform safety functions must conform to an increasing set of regulatory requirements. In the nuclear energy domain, a licensee must therefore demonstrate that his system meets all regulatory requirements of a regulator. These requirements can be contained in regulatory documents, in guides, standards and even in tacit knowledge [22] acquired from anterior projects in the past. This lays applicants with a huge and increasing amount of documents and information which is mostly not formalized.

This work takes its root on Instrumentation and Control (I&C) systems in nuclear power plants. I&C systems include instrumentation to monitor physical conditions in the plant (e.g., temperature, pressure, or radiation), redundant systems to deal with accidental conditions (safety systems) and all the equipment for human operators to control the behavior of the plant. While digital components are replacing most of the older conventional devices in I&C systems, confidence in digital technologies remains low. Consequently, regulatory practice evolves and new standards appear regularly while domain expertise is heavily involved for certification.

The major issue for licensees who must assess conformance to all regulatory requirements is the lack of traceability between all regulations, practices accepted by one regulator, standards and technical requirements. Consequently, licensees and regulators rely more and more on human expertise for assessment, increasing the amount of scattered tacit or not formalized knowledge in the process. If operators experts have a quite precise knowledge of the regulatory context in their country, this knowledge is not capitalized yet. As operators tend to build plants in foreign countries, they have to face new regulations or different practices upon a similar regulation. In this new context, they mostly have to re-qualify their system from scratch to fit targeted country regulations.

This paper is an initial proposal towards the identification of major themes in the corpus, around which we can establish traceability links. A theme in our case is a concern for one of the experts involved in the assessment process.

We address the following research questions:

- What are these requirements and how are they related?
- Can we determine the different themes included in such documents and localize the area where these themes are addressed in order to reduce the problem space and ease the traceability analysis?

Recently, Gotel and Morris [11] highlight the challenges specific to requirements traceability and illustrate how existing practices can be leveraged to tackle this challenge. Such analysis can be performed through the efficient use of information retrieval (IR) methods which may be able to raise valuable information from textual units contained in a regulatory corpus, as IR has proved to work in an efficient way for more requirements traceability [4][12][8].

In this paper, we first present the industrial context as well as an illustrative example of a regulatory concern flow in two contexts. We then introduce an approach using natural language processing and information retrieval techniques to define and retrieve themes in order to have a global but more precise view of a theme.

The remainder of the paper is organized as follows. Section II details the first contribution of the paper: a synthetic overview of the regulatory requirements landscape in the nuclear domain and the traceability challenges it encompasses. Section III presents a concrete illustrative example of regulatory requirements in motion. Sections IV and V discuss definitions and present the approach as well as challenges related. In Sections VI and VII, we discuss related work and present some perspectives for future work.

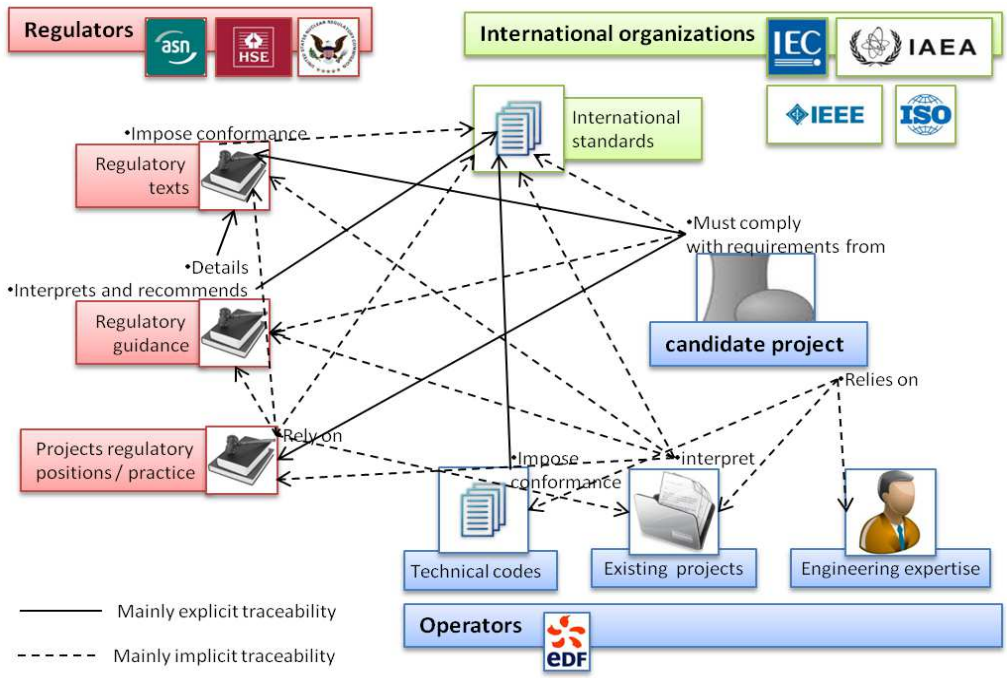

Figure 1 Overview of the nuclear regulatory landscape

II. QUALIFICATION OF SAFETY SYSTEMS AND NATIONAL PRACTICES REGARDING I&C SYSTEMS

In a quite recent history, answering to a nuclear industry motto: "to cope with complex safety problems, the simpler the solution is, the better the solution is", nuclear industry was used to utilize relays and conventional (not digital) technologies, which were simple enough to be used and qualified for complex and critical safety functions.

Digital systems have now become essential in all industries and these conventional components are not available anymore in the market and less and less specified for nuclear industry sole usage like COTS (Commercial Off-The-Shelf). Unfortunately, it represents a monumental effort to try to demonstrate, if feasible, the complete absence of error into these digital systems. The situation becomes worse while relating to some famous failures due to software during the last decades.

Based on their experience acquired from past or recent projects, regulators of each country have built a unique and specific practice related to nuclear energy and safety concerns. This section provides an overview of the regulatory requirements corpus related to safety in nuclear I&C systems. We focus on all the links that must be established in order to certify a system.

A. *Operators and regulations*

Figure 1 gives an overview of the different kind of documents and actors involved in the safety assessment process for a candidate plant project. We detail this figure and illustrate it within the scope of digital I&C systems.

When licensees, like EDF (Electricité de France), plan a project (realization of a new power plant, substitution of obsolete technologies in existing plants, renewal of an exploitation license), they rely on their experience acquired on past projects or take into account other existing projects. They may have issued technical codes to ease reusability along their different projects. They also rely on their engineering expertise to cope with complex emerging technical issues when innovation is required.

The proposed solution must comply with regulatory requirements. These requirements or recommendations are expressed in multiple documents: legal documents issued by national authorities; standards, issued by international organizations; regulatory practices, which arise from specific questions from regulators and following discussions. These different types of requirements, shown at the left and top of Figure 1, are detailed in the following.

Regulatory requirements are complete in the sense that there are no others (even if you should consider them as incomplete). They are ambiguous [14], not clear and unverifiable. Finally, there is no way (within the scope of qualification) to change and improve them. Thus, these requirements are far from the usual separation between functional/non functional requirements and they are not concerned with requirements quality where the objectives are more to produce complete, verifiable, precise requirements or to try to reach this final state.

B. *Different kind of regulatory texts*

1) *Regulatory texts with regulatory requirements*

Regulatory texts issued by public authority, express very high level requirements, principles or objectives related to people's life and environment protection, applicants responsibilities and duties. These texts do not provide guidance to achieve these requirements.

In France, such documents and requirements are collected in the *Basic safety rules* documents (RFS II.4.1.a related to software, issued in French). In the USA, they are expressed through the Code of Federal Regulations *10CFR50* and its appendices. In the UK, the requirements are collected in the *Safety Assessment Principles* (SAPs).

2) Regulatory guidance

Regulatory guides describe the regulator's position and what he considers as an acceptable approach. These guides, endorse (or not) parts of standards and may provide interpretations of some specific parts.

In France, there is no such document available. In the USA, the Nuclear Regulatory Commission (NRC) publishes regulatory guides such as the Regulatory Guide 1.168 for *Verification validation, release and audit for digital computer software used in safety systems*. In the UK, one can find the Technical Assessment Guides (TAGs), for example the TAG 003 titled *safety systems* and 046 titled *Computer-based safety systems.*

3) Regulatory positions and practice

During projects submissions, realizations, operations, maintenance, licensees still have to deal with regulators and issue documentations related to a specific project or installation. It can be the case for example for the renewal of an obsolete I&C system which raises a problem of qualification of a new device.

This leads to regulatory positions while accepting or refusing propositions (for instance, the authorization of operation for ten more years for one reactor in France) or requiring improvements on specific topics. This is the most explicit highlight of the regulatory practice.

C. International standards and practice

International standards are state of the art propositions covering specific domains. It is important to notice that the requirements and recommendations in these standards are meant to be applied in a voluntary way, except when a regulator imposes or recommends its application. At this moment, standards requirements are considered as regulatory requirements. One other important aspect to consider is that different standards may exist to deal with the same subject. In Europe, nuclear actors mainly follow the IEC/IAEA corpus whereas in the US, IEEE/ISO standards are applied. These two corpora have been written independently from each other.

Standards include external traceability links to others documents ("normative references") and each document possesses vocabulary that is merely defined in "definitions" and "Symbols and abbreviations" sections. Elements of standards are contained into sections related to a particular concern. These elements may reference both internal and external other elements and documents.

III. TRACKING A CONCERN IN NUCLEAR REGULATIONS

The following example is a manual analysis so it is impossible to evaluate the completeness of the coverage of the topic. Yet, it illustrates the complexity of the regulatory landscape depicted in Section II.

Considering specific purpose analysis such as V&V regulatory requirements in safety systems in different countries, one should initially think that these requirements are close enough to be compared. Let's take an example of what we have to face at a very high level and refine it to the normative level in two different contexts: France and USA. Provided examples are very short excerpts from the documents to illustrate both, kind of sentences and different concerns at different abstraction levels.

A. At the regulatory level

In France, in the RFS (basic safety rule) II.4.1.a (2000), the requirements or principles are written in French. About the concern Verification and Validation, Figure 2 proposes a translation.

In the USA, we shall consider the 10CFR50 and in particular following excerpt in Figure 4.

At this level, we can agree that there are mainly common points regarding verification and validation even if it is not mentioned in the US regulation (apart from the word "tested". In France, independent V&V is already explicit. Fitness to specification (validation) is present. In all of them, quality assurance programs are mentioned. The notion of compliance with standards is expressed everywhere with more or less importance. Software safety life cycle is approached using different terms using enumeration of activities in the US, fitness to specification, V&V methods in France). We also observe the emergence of different level of application of standards as acceptable approaches (France, US, UK), best in process and applicability (US) and mandatory items (US).

Requirements for software in category 1E programmed systems
Functions
Reliability
Reliability is addressed within qualitative perspectives
Ea 2.1 *Software design and documentation shall allow performing verification and validation methods in order to demonstrate … An acceptable practice, related to methods and techniques of verification, is described in chapters 6 (verification) and 7 (software/component integration) of the IEC 60880 publication (1986)… Similarly, simulation is an acceptable technique for the validation of the executable program, especially for time performances. This technique can be combined with prescriptions of chapter 8 of IEC60880 publication (1986).*

Figure 2 V&V in French regulatory text

Par55a(a)(1): Codes and Standards

(a) Quality standards, ASME Codes and IEEE standards, and alternatives.

(1) Structures, systems, and components must be designed, fabricated, erected, constructed, tested, and inspected to quality standards commensurate with the importance of the safety function to be performed. …

(h) Protection and safety systems.

(2) Protection systems. For nuclear power plants … must meet the requirements stated in either IEEE Std. 279 … or in IEEE Std. 603-1991 …

(3) Safety systems. Applications … must meet the requirements for safety systems in IEEE Std. 603–1991 and the correction sheet dated January 30, 1995.

Appendix A to Part 50--General Design Criteria for Nuclear Power Plants

I. Overall Requirements

Criterion 1— Quality standards and records. Structures, systems, and components important to safety shall be designed, fabricated, erected, and tested to quality standards commensurate with the importance of the safety functions to be performed. …

Figure 4 US 10CFR50 regulation

B. At the regulatory guidance level

There is no document at this level in France. Nevertheless, the RFS explicitly mention that use of Chapter 6, 7 and 8 of the IEC60880 (1986) are acceptable practices for software V&V of category 1E systems. As the French safety authority closely works with EDF, it has endorsed the RCC (Rules for Design and Construction) series issued by EDF (considered as a technical operator code in Figure 1). In particular, RCC-E (for electrical devices) requires conformance with several international standards such as IEC60880, IEC62138, etc. depending on the safety function category performed by the software.

In the US, it is described partially into the regulatory guide 1.168 (excerpt in Figure 3) that will later lead us to the analysis of the IEEE standard 1012. This guide is a rather small document (only 11 pages) with backward traceability to 10CFR50.

*This regulatory guide **endorses IEEE Std 1012-1998**, "IEEE Standard for Software Verification and Validation," and IEEE Std 1028-1997, "IEEE Standard for Software Reviews and Audits." IEEE Std 1012-1998, with the exceptions stated in the Regulatory Position, describes a method acceptable to the NRC staff for complying with parts of the NRC's …*

C. REGULATORY POSITION

IEEE Std 1012-1998, "IEEE Standard for Software Verification and Validation," provides methods that are acceptable to the NRC staff for meeting the requirements of 10 CFR Part 50 as they apply to the verification and validation of safety system software, subject to the exceptions listed in these Regulatory Positions. …

The annexes to IEEE Std 1012-1998 and IEEE Std 1028-1997 contain information that may be useful, but the information in these annexes should not be viewed as the only possible solution or method. …

Figure 3 US IEEE Regulatory Guide 1.168

These sentences confirm the traceability link between this guide and the IEEE standards 1012 and 1028 and that interpretation of some fragments will appear. In particular, annexes or pieces of standards may or not be endorsed by the regulator. They define the set of requirements which will be applicable while desiring to comply with the standard.

C. At the normative level

From this moment on, whereas previous documents were freely accessible, standards and more technical documents become proprietary and less easily accessible. Beyond the three tracks followed above, the next step finally leaves us with two documents from the IEC and IEEE community. If both IEC60880 and IEEE1012 deal with software validation and verification, the chosen perspective of description is rather different.

IEC 60880 (chapter 8) deals with:

1. Independence of the verification;
2. Verification plan;
3. Design verification;
4. Implementation verification (with both general purpose and application-oriented languages and respective test reports);
5. Configuration of pre-developed software.

IEEE 1012 deals with:

1. Software V&V processes: management, acquisition, supply, development, operation, maintenance;
2. Software V&V reporting, administration and documentation;
3. Detailing a software V&V plan outline.

Each of these processes is detailed through several tasks, required inputs and outputs and including some specific traceability/interface/risk/hazard/security analysis.

Standards contents though do not express the same requirements about the same activity. IEC 60880 expresses objectives to reach whereas IEEE 1012 details activities to perform to reach objectives.

D. Synthesis

We just manually performed a track retrieval experiment for the theme "software verification" in French and US corpora. Still, we can observe major differences as documents are written within different objectives at all levels of the regulation hierarchy. This difference is the most explicit at the standard level. IEC60880 depicts achievement requirements whereas IEEE1012 depicts process requirements as it is not nuclear specific. Thus, it could seem that we are comparing apples and oranges but, yet, it provides some useful information. On the one hand, being IEC60880 compliant for this theme does not provide a straightforward IEEE1012 compliance as software verification is not assessed using the same criteria. On the

other hand, they share the same principles and final objectives and may complement each other.

E. Standard and practices gaps

More generally, there is a gap between the IEC corpus, which is specifically written by the IEC subcommittee SC45-A and that issues nuclear specific to nuclear industry and IEEE standards which are not always nuclear specific, for instance, IEEE1012 deals with general software Validation and Verification activities.

We can illustrate this gap by comparing concerns of the different used standards in the same digital I&C context but in two different countries: France and USA.

In France, we can cite the 8 following standards that cover a large scope of digital I&C systems (Complete titles are all prefixed with "*Nuclear power plants – Instrumentation and control important to safety*".

- *IEC 60880-2006 Software Aspects for Computer-Based Systems Performing Category A Functions*
- *IEC 60987-2007 Hardware Design Requirements for Computer-Based Systems*
- *IEC 61226-2009 Classification of Instrumentation and Control Functions*
- *IEC 61500-2009 Data Communication in Systems Performing Category A Functions*
- *IEC 61513-2011 Nuclear power plants – Instrumentation and control important to safety – General Requirements for Systems*
- *IEC 62138-2004 Software Aspects for Computer-based Systems Performing Category B or C Functions*
- *IEC 62340-2007– Requirements for Coping with Common Cause Failure (CCF)*
- *IEC 62566-2011 Development of HDL-programmed Integrated Circuits for Systems Performing Category A Functions*

By the same time, the NRC, The US regulator proposes a clear snapshot of its regulatory context by endorsing explicitly parts of standards. For I&C systems, these standards are:

- *IEEE 338-1987 Criteria for the Periodic Surveillance Testing of Nuclear Power Generation Station Safety System*
- *IEEE 7-4.3.2-2003 Standard Criteria for Digital Computers in Safety Systems*
- *603-1998 Standard Criteria for Safety Systems for Nuclear Generation Stations*
- *IEEE 1028-1997 Standard for Software Reviews and Audits*
- *IEEE 1012-1998 Standard for Software Validation and Verification*
- *IEEE 828-2005 Standard for Software Configuration Management Plans*
- *IEEE 829-1998 Standard for Software Test Documentation*
- *IEEE 1008-1993 Standard for Software Unit Testing*
- *IEEE 830-1998 Recommended Practices for Software Requirements Specifications*
- *IEEE 1074-1995 Standard for Developing Software Lifecycle Processes*

Johnson [13] attempted to align these two corpora. Yet his alignments were made only at the document level and were published in 2001. Completing a full alignment requires determining common concerns to analyze and then comparing documents contents, from definitions to requirements.

We can observe that both collections have not the same approach neither the same requirements against safety systems of nuclear power plants. The French collection deals with very specific concerns (common cause failure, data communication, etc.) whereas the US collection deals with steps of system's lifecycle. In these contexts, the US regulator endorses both general and nuclear specific standards whereas the French standard collection is nuclear specific as to cope with nuclear specific issues. It describes two different regulatory practices with their own particularities.

It is clear that the different requirements cannot be merged easily and that there is no one-to-one traceability links inside a corpus or mapping within the same collection or across different corpora. We need to split these regulatory corpora into smaller and organized units of concerns to be able to better capitalize the regulatory domain knowledge and also perform computable analyses such as impact, similarity or coverage analyses in smaller but more relevant sets.

IV. THEMES TO NARROW RESEARCH SPACE IN REGULATORY CORPORA

A. Collecting themes' traces through the entire corpus

As illustrated previously, the basic requirements to comply with are contained into the regulatory texts and the regulatory guidance. Without this minimal subset, an applicant cannot apply to any project. Yet, the detailed knowledge relies not only on these documents, but on previous assessment processes done throughout history: accepted practice on past and recent projects. All these documents do not provide the same level of requirements or recommendations but each of them is necessary to understand the global qualification process.

As shown in Figure 1, most of the links between these documents are implicit links. There are several reasons for that. First, regulations shall not be ad hoc decisions and shall persist over the years. Standards documents result from stakeholders' negotiations. As a consequence, they are ambiguous in both unintended and intended ways [1]. Second, regulations do not evolve as quickly as technology

as illustrated in our previous work [20]. Third, there is a chronological variability dimension around the regulatory documentation. On the one hand, guidance on a topic cannot be written before the regulatory text it explains. On the other hand, regulatory texts are not automatically updated with each domain modification. For example, the Software basic safety rule in France, issued in 2000, has not been updated yet to consider the current practice, which includes many more standards than IEC60880, related to software aspects. This hinders forward and backward traceability [10] and tends to increase the list of implicit traceability links and implicit cross-references [16].

Yet, we cannot retrieve a trace between a complete regulatory text and a complete standard. Instead, it is necessary to extract coherent subsets from the standard that can be related to subsets of regulations. In the following, we call such a coherent subset in a corpus a *theme track*. In the rest of the paper we focus on the set of IEC standards related to safety.

We focus on standards rather than regulatory texts, because regulation, guidance and positions are specific to the countries in which they are published. They have a very high level of abstraction, which leads to a lot of interpretation upon the same document. Yet, regulators also discuss around standards, and since they capture the state of the art practices, they are more shared. Standards also represent the most precise layer of regulatory documents (when imposed by a regulator), just before operators' documents. They represent a good balance between abstract regulatory documents and operators' technical documents.

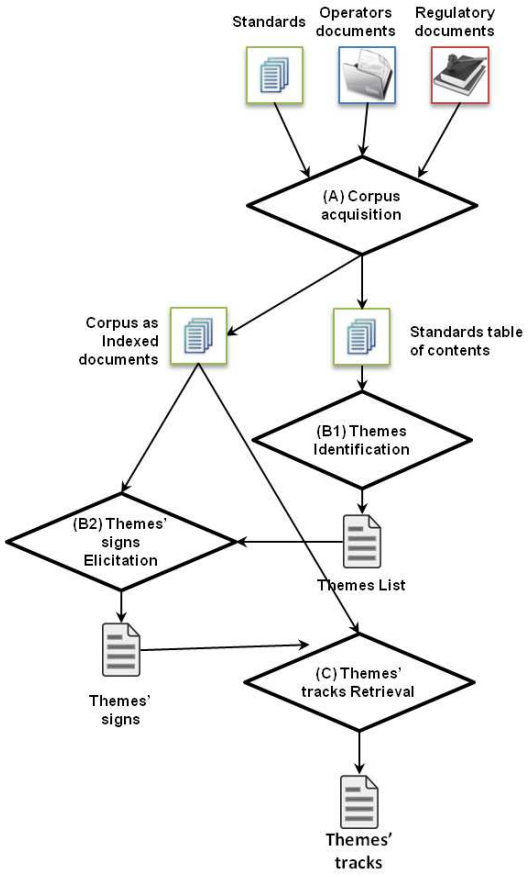

Figure 5 Defining, Building and Retrieving Themes

B. Principles

In this section, we propose definitions and the theoretical approach. We also expose important concepts of information retrieval.

1) Definitions

1. A *theme* is a concern within a corpus (e.g., "common cause failure", "maintenance", etc.). It is defined by *theme signs* and represents a viewpoint on a corpus. It contains *theme tracks* related to this topic.
2. *Theme signs (or signature)* are defined by Gotel and Morris [11] as an *"identifying mark made by, or associated with a particular purpose, an animate or inanimate object"*. In our case, we consider the terms that are specifically related to a theme as the signature elements that identify a theme.
3. *Theme tracks* are the collection of textual excerpts of the corpus that are relevant to a given theme. A textual excerpt may belong to several themes.

Figure 5 displays the analysis flow we want to follow. The first step consists in gathering the different documents in a computable way (A) and acquiring an initial corpus model. From the extracted table of contents of the documents to add, we determine the different themes, which are consistent with the domain (in our case, digital I&C systems) (B1). Once the theme list is determined, the second step is about building each theme's signature by collecting its signs (B2). These signature elements will allow detecting the theme all over the full corpus and across different corpora. Provided each theme's signature, the last step consists in retrieving the areas of the corpus related to a theme (C) where alignment or impact analyses can be performed in an easier way.

2) Information retrieval to support the Theme approach

a) Document

The basic concept for information retrieval techniques is the "document". These documents contain different information named after "fields" such as document's authors, title, content, URL, etc. Indexing a document is the action of filling these information fields from the document in an efficient way. Searching into an index is related to a specific field of the index.

b) "Stop" words and stemming

Stop words are common words of the language such as "*the*", "*any*", that hold no particular meaning. There exist lists of stop words for many languages. Stop words filtering removes these words from the analyzed document.

Stemming transforms a word in its root form. Its aim is to cluster different forms of a word that would have been

disassociated otherwise. For instance, the title *"specific requirements related to blank integrated circuits"*, after these two steps will be: "*specif requir relat blank integr circuit*".

c) Term frequency (TF), Inverse document frequency (IDF) and TF-IDF weight

The term count is the number of occurrences of a given term in a given document. This count is usually normalized to prevent a bias towards longer documents and the resulting term frequency TF (t,d) gives a measure of the importance of the term *t* within the particular document *d.*

The inverse document frequency is a measure of the general importance of a term. Common terms have low IDF scores and the rarer they are, the higher they score.

It is computed as $idf\ (t) = log\ N/df_t$ where N is the total number of documents in the corpus and df_t the number of documents that contains term t.

The tf-idf weight [19] is a numerical statistic which reflects how important a term is relatively to a document in a corpus.

It is computed as $tf\text{-}idf(t,d,D) = tf(t,d)*idf(t)$.

There exist different customizations of the basic formula, favoring recall, weighting terms, etc. [5][6].

V. CHALLENGES FOR THEME TRACKS RETRIEVAL IN REGULATORY CORPORA

In this section we present challenges for steps of the Figure 5 process. Section V.A is related to the corpus acquisition step. Section V.B discusses themes identification. V.C addresses the third step and theme retrieval in the corpus.

A. Corpus acquisition

As mentioned in Section II.A, we manipulate documents that are imposed, with a limited access to the document sources. The way these documents are written is very variable and does not respect any general pattern over the different years of publication. Tables and figures are particularly challenging for automated analysis, since their structure is lost in the logical text stream.

This step can be performed in two ways: Manual slicing and indexing of the documents collections, or an automatic parsing. Manual slicing allows straight forward indexing of the collection and direct searching.

(Semi)Automatic acquisition requires developing a generic parser to analyze and dispatch the different information of each document into a model, yet it would represent a time consuming effort. Consequently, it has a significant impact while trying to organize and add properties as it also means to provide a metamodel that represents the domain to capitalize as well as rules to transform textual fragments as rich model elements.

Evaluation of such approach is done with respect to the conformance of the generated data with the domain metamodel and expert validation.

B. Themes identification using standards table of contents

1) Determining themes using statistical measures

Computing term-frequency analysis over the corpus headlines aims at recovering top emerging words and grouping similar inputs in order to identify themes. Yet it introduces several different issues of polysemy and synonymy we detail later. Such evaluation requires an empirical validation and a subsequent work toward formalizing the domain vocabulary utilizing regulatory documents definitions, leveraging domain expertise, etc.

2) Determining themes using a clustering algorithm

Using a clustering algorithm to build clusters of similar documents offers traceability while exploiting the result. Because textual fragments may belong to several themes (see definition), we need an algorithm that allows overlapping clusters and multi-words cluster naming as documents may belong to several different themes and multi-words themes such as "common cause failure", which is a specific concern, should be able to be built. Evaluation of the relevance of determined themes can be done using headlines' coverage analysis.

3) Synonymy, polysemy, ambiguity

Synonymy (multiple words for one meaning) and polysemy (multiple meanings for one word) are issues identified very early in natural language analysis toward computable analysis. Within a similar separation of concerns purpose, it has been described in [1].

WordNet dictionary [23] is a general language dictionary and may not particularly fit technical domains. For instance, the noun "design" has the following synonyms: *{aim, blueprint, conception, contrive, designing, excogitation, figure, innovation, intent, intention, invention, pattern, plan, project, purpose}*. However, it is necessary to align for example "V&V" in the usual systems engineering vocabulary vs "independent confidence building" used in UK, and analyze whether these expressions are related to each other.

Coping with this concern requires a thorough analysis of the domain vocabulary and practices. However, regulatory documents usually use a well defined vocabulary and hold a definition section to disambiguate terms. That may help to tackle this concern.

4) One word tokens vs multiwords tokens

"C*ommon cause failure*" is a very specific concept and shall be represented as-is instead of separated tokens: "*common*", "*cause*" and "*failure*", where failure may be another topic itself. Similarly, more general concepts such as "*requirements specification*" should be analyzed as separated tokens and grouped ones. Our initial intuition is to regroup them as such concept should be differentiated from the general "*requirements*" and "*specification*" concepts.

5) Establishing theme signs

Gotel and Morris [11] established traceability analysis in terms of signs belonging to individuals and tracks related to them. It fits to the concept of theme signs and tracks.

Unfortunately, finding indicators that clearly identify a theme (specific terms and occurrences of these terms) is a clear issue when documents use a very precise lexicon that prevents the emergence of additional indicators. This issue has been addressed by Gibiec et al. [8] and needs further investigation in our context. This concern is also impacted by the previous concern described in the previous section: synonymy and polysemy. Establishing theme signs and retrieving themes are closely related operations as both steps operate on the corpus.

For this step, we focus on standards rather than regulatory texts and use them as learning sets. Regulation, guidance and positions are specific to the countries in which they are published. They have a very high level of abstraction, which leads to a lot of interpretation upon the same document. Yet, regulators also discuss around standards, and since they capture the state of the art practices, they are more shared. Standards also represent the most precise layer of regulatory documents (when imposed by a regulator), just before operators' documents, which also may also reuse standards vocabulary. They represent a good balance between abstract regulatory documents and operators' technical documents.

C. *Retrieving theme tracks in the corpus*

In previous step, themes are identified on the basis of standards table of contents, themes signature are determined using a learning set made of standards. Now, we consider section's whole content to retrieve themes in the whole corpus.

Evaluation of the retrieved themes' tracks shall be performed using a recall/precision evaluation. Yet, in traceability analysis, search approaches usually record good recall and poor precision [4] and though generate a lot of candidate link. To narrow this, it is usual to set a threshold (cutoff value) below which scores are not taken into account, considered as too poorly related to the initial query. Documents are ordered according to their score given by the retrieval operation. Usual IR approaches for traceability analysis favor recall over precision. Dealing with false positives is easier to deal with than omission errors (false negatives) [12]. However, providing a 100% recall score may be irrelevant.

Once themes tracks are gathered, it is possible to organize them with respect to relationships defined in the domain metamodel such as traceability links between different fragments that one could want to highlight or forward traceability toward architecture elements, etc. This could be such as dependencies we defined in a previous work [20], where we defined refinements and interactions: allocation, justification, qualification links for traceability aspect around the system lifecycle or (total/partial) equivalence, conflicts, coverage, requires, reference links to define relationships between documents. Other examples of relationships may consist in those defined by Maxell et al. [16] or dependencies defined by Zhang et al. [24].

VI. RELATED WORKS

Natural language processing (NLP) and information retrieval approaches have been previously used for Requirements Analysis. At the system's scale, it has been pioneered by Sawyer et al. [21] within the REVERE project and tool while having initial results in detection of roles and "shall"/"should" to distinguish between requirements types. Kiyavitskaya et al. [15] use GaiusT to extract rights, obligations, on both HIPAA (Health Insurance Portability and Accountability Act) and equivalent Italian regulations. It is not based upon a term-frequency analysis but relies on text decomposition in a parse tree conforming to a structured grammar and fragments annotations. More recently, Cleland et al. [4][8] use NLP techniques to trace regulatory requirements from HIPAA in several software applications. In their subsequent work, they combine NLP with clustering and association rules to recommend features [6]. Though the followed process is very similar, it is worth noticing that we do have differences in some specific experimental choices: We work on very different documents, industrial standards, with different constraints. We sliced our documents in a way that keeps the document structure instead of arbitrary chunked fragments. The Lingo clustering algorithm we used allows overlapping, which reflects scattering of themes throughout the corpus.

About regulatory requirements and compliance concerns, extensive studies had been done in healthcare domain and, particularly around HIPAA. In [17], production rules are developed to translate regulatory texts into production rules to represent a formalized form of legal knowledge and address ambiguity. In [3], the authors use semantic parameterization to derive rights and obligations from HIPAA and compare different stakeholders' interpretations. In [9], specific laws statements of multiple jurisdictions about data breach are refined using a requirement specification language. Statements are then neighbored and similar ones are organized to identify gaps, conflicts and try to reconcile them. In [16], the authors focus on explicit external cross-reference links and propose a legal cross-reference taxonomy to formalize these relationships. In [7], the authors use User Requirements Notation (URN), a combination of NFR and i* frameworks and use-case maps, to model both the Personal Health Information Privacy Act and a hospital business process and assess its compliance against the privacy law.

These work concentrate on the law level and explicit traceability links while we expect to follow a flow that covers multiple levels of documentation. The proposed taxonomy in [16] is close to dependencies we defined in [20]. We do not address directly the compliance issue here, though, assisting experts to retrieve implicit links in a shorter than initial problem space may represent a way to achieve it.

Related to software standards analysis for qualification purposes, [25] and [18] propose model-driven engineering approaches and use UML profiles to address specifically the

DO-178B and IEC61508 standards. DO-178B is a standard dedicated to software aspects in the aerospace domain. The proposition aimed to maintain traceability from requirements to design to code that we do not address here. In [18], the authors gather concepts from the standard and build a conceptual model of the IEC61508 standard. As a consequence, the proposition remains specific to IEC61058 and was targeted to address the safety evidence question, whereas we need a more general framework.

VII. CONCLUSION

In this paper, we presented first, the nuclear regulatory requirements landscape, which is a fairly new domain for requirements analysis. This domain is complex because of the variety of documents one has to handle; the number of requirements they contain; their high level of abstraction and ambiguity, and their implicit and complex relationships that are an issue for both requirements analysis and traceability analysis concerns. We addressed the question of narrowing this problem space by clustering it around the concept of themes. We provided our definition for a theme and proposed an approach using natural language processing and information retrieval techniques to define and retrieve a theme into a corpus of documents. We discussed different challenges over the approach. In future work, we plan to address the different challenges we discussed in Section V. We actually also work on a domain specific modeling language in order to organize information contained into a regulatory corpus [20] and provide richer traceability information.

ACKNOWLEDGMENT

This work is partially supported by the EU FP7-ICT-2009.1.4 Project N° 256980, NESSoS: Network of Excellence on Engineering Secure Future Internet Software Services and Systems.

REFERENCES

[1] E. L. A. Baniassad, and S. Clarke, "Theme: An Approach for Aspect-Oriented Analysis and Design", In (ICSE'2004), Edinburgh, Scotland, pp. 158-167, 2004.

[2] T. D. Breaux, and A. I. Antón, "A Systematic Method for Acquiring Regulatory Requirements: A Frame-Based Approach", In (RHAS-6), Delhi, India, 2007.

[3] T. D. Breaux, M. W. Vail, and A. I. Antón, "Towards Compliance: Extracting Rights and Obligations to Align Requirements with Regulations", In (RE'06), Minneapolis, USA, pp. 49-58, 2006.

[4] J. Cleland-Huang , A. Czauderna, M. Gibiek, and J. Emenecker, "A machine learning approach for tracing regulatory codes to product specific requirements", In (ICSE'2010), Cape Town, South Africa, pp. 155-164, 2010.

[5] C. Duan, and J. Cleland-Huang, "Clustering support for automated tracing", In (ASE'07), Atlanta, USA, pp. 244-253, 2007.

[6] Dumitru, H., M. Gibiec, N. Hariri, J. Cleland-Huang, B. Mobasher, C. Castro-Herrera, and M. Mirakhorli. "On-demand feature recommendations derived from mining public product descriptions", In (ICSE'2011), Honolulu, USA, pp. 181-190, 2011.

[7] S. Ghanavati, D. Amyot, and L. Peyton, "Towards a Framework for Tracking Legal Compliance in Healthcare", In (CAiSE'07), Trondheim, Norway, pp. 218-232, 2007.

[8] M. Gibiec, A. Czauderna, and J. Cleland-Huang, "Towards mining replacement queries for hard-to-retrieve traces", In (ASE'2010), Antwerp, Belgium, pp. 245-254, 2010.

[9] D. G. Gordon, and T. D. Breaux. "Comparing Requirements from Multiple Jurisdictions", In (RELAW'2011), Trento, Italy, pp. 43-49, 2011.

[10] O. Gotel, and A. Finkelstein, "An Analysis of the Requirements Traceability Problem", In (RE'94), Colorado Springs, USA, pp. 94-101, 1994.

[11] O. Gotel, and S. J. Morris, "Out of the labyrinth: Leveraging other disciplines for requirements traceability", In (RE'11), Trento, Italy, pp. 121-130, 2011.

[12] J. H. Hayes, A. Dekhtyar, and S. K. Sundaram, "Advancing candidate link generation for requirements tracing: the study of methods", In IEEE TSE, vol. 32(1), pp 4-19, 2006.

[13] G. Johnson, "Comparison of IEC and IEEE standards for computer-based control systems important to safety", Nuclear Science Symposium Conference Record, (2001).

[14] E. Kamsties. "Understanding Ambiguity in Requirements Engineering." In Aybüke Aurum & Claes Wohlin, editors, Engineering and Managing Software Requirements, chapter 11, pp 245-266. Springer, 2005.

[15] 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 (ER'08), Barcelona, Spain, pp. 154-168, 2008.

[16] J. Maxwell, A. I. Antón, and P. Swire, "A Legal Cross-References Taxonomy for Identifying Conflicting Software Requirements", In (RE'11), Trento, Italy, pp. 197-206, 2011.

[17] J. Maxwell, and A. I. Antón, "Developing Production Rule Models to Aid in Acquiring Requirements from Legal Texts", In (RE'09), Atlanta, USA, pp. 101-110, 2009.

[18] R. K. Panesar-Walawege, M. Sabetzadeh, and L. Briand, "A Model-Driven Engineering Approach to Support the Verification of Compliance to Safety Standards". In (ISSRE'11), Hiroshima, Japan, pp. 30-39, 2011.

[19] G. Salton, "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer", Addison-Wesley, 1989.

[20] N. Sannier, B. Baudry, and T. Nguyen, "Formalizing standards and regulations variability in longlife projects. A challenge for Model-driven engineering", In (MoDRE'2011), Trento, Italy, pp. 64-73, 2011.

[21] P. Sawyer, P. Rayson, and R. Garside, "REVERE: support for requirements synthesis from documents", Information Systems Frontiers Journal. Volume 4, issue 3, Kluwer, Netherlands, pp. 343-353, 2000.

[22] P. Sawyer, V. Gervasi, and B. Nuseibeh, "Unknown knowns: Tacit knowledge in requirements engineering", In (RE'11), Trento, Italy, p. 329, 2011.

[23] Wordnet, Princeton University, "About WordNet", http://wordnet.princeton.edu , 2010.

[24] W. Zhang , H. Mei, and H. Zhao, "A Feature-Oriented Approach to Modeling Requirements Dependencies", In (RE'05), Paris, France, pp. 273-284, 2005.

[25] G. Zoughbi, L. Briand, and Y. Labiche, "Modeling safety and airworthiness (RTCA DO-178B) information: conceptual model and UML profile", Software and System Modeling 10(3): 337-367, 2011.