Classification, Structuring, and Assessment of Evidence For Safety

A Systematic Literature Review

Sunil Nair[1], Jose Luis de la Vara[1]
[1]Certus Centre for Software V&V
Simula Research Laboratory
P.O. Box 134, 1325 Lysaker, Norway
{sunil, jdelavara}@simula.no

Mehrdad Sabetzadeh[2], and Lionel Briand[2]
[2]SnT Centre for Security, Reliability and Trust
University of Luxembourg
6, rue Coudenhove-Kalergi L-1359 Luxembourg
{mehrdad.sabetzadeh, lionel.briand}@uni.lu

Abstract— **Safety assurance and certification are amongst the most expensive and time-consuming tasks in the development of safety-critical systems. Demonstration of compliance with safety standards involves providing evidence that the standards' safety criteria are met. To handle large collections of evidence effectively, safety professionals need knowledge of how to classify different types of evidence, how to structure the evidence, and how to assess it. This paper takes a step towards developing such a body of knowledge by conducting a Systematic Literature Review (SLR). Specifically, the SLR identifies and classifies the information and artefacts considered as evidence for safety, examines existing techniques for evidence structuring and assessment, and summarizes the challenges noted in the literature in relation to safety evidence. The paper, to our knowledge, is the first systematic review on the topic of safety evidence. The results we present are particularly relevant to practitioners seeking to better understand the evidence requirements for safety certification, as well as to researchers conducting research in this area.**

Keywords-safety-critical systems; safety standards; safety compliance; safety certification; safety evidence; systematic literature review.

I. Introduction

Most critical systems in domains such as avionics, railway, and automotive are subject to a safety assurance process as a way to ensure that these systems do not pose undue risks to people, property, or the environment. The most common type of assurance is safety certification [31], whose goal is to provide confidence that a system is deemed safe by an independent licensing or regulatory body.

Underlying any kind of safety assurance process is a set of industry-accepted criteria, typically available as standards that need to be complied with. Notable examples of safety standards are IEC61508 [17] for a broad class of systems, DO-178C for avionics [9], the CENELEC standards for railway [45], and ISO26262 for the automotive sector [10].

Demonstrating compliance with a standard involves the collection of evidence that shows that the safety criteria envisaged by the standard are met [25]. Examples of evidence are hazard analyses, design specifications, and test results. To this end, an important task for the system suppliers and the assessors is to elaborate and agree on what evidence is exactly made up of in a given context, how the evidence should be structured and assessed in order to verify that the safety criteria of interest have been met.

Without a clear understanding of the evidence needs, two main problems may arise. First, the supplier may fail to record critical details during system development that the assessor will need, in turn requiring the supplier to reconstruct the missing evidence after-the-fact. This can be both expensive and laborious [47][53]. Second, the assessor may find it hard to develop enough confidence in the system undergoing assessment without having agreed to the evidence requirements first [13].

In addition to defining precisely the information that the evidence should cover, attention needs to be paid to how this information is organized. Particularly, for large-scale systems, if the evidence is not structured properly, its sheer volume and complexity can jeopardize the clarity of the high-level safety arguments [55].

Finally, the supplier and the assessor need to establish an a priori understanding of how the collected evidence will be assessed. Without this understanding, the supplier may be unable to determine the importance of the different aspects of evidence. As a result, they may risk spending a disproportionate level of effort on evidence aspects with marginal contributions to achieving compliance, while not spending enough effort on evidence aspects that are critical for compliance [13].

The main objective of this paper is to synthesize the existing knowledge in the academic literature about safety evidence, concentrating on the three facets outlined above, namely, the information that constitutes evidence, structuring of evidence, and evidence assessment. We achieve this objective by means of a *Systematic Literature Review* (SLR) – a documented and repeatable process through which the literature on a given subject is examined and the current state of knowledge is recorded [29]. The main advantage of a SLR, when compared to ad hoc search, is that it provides a higher degree of confidence about covering the relevant literature and thus minimizes subjectivity and bias.

Our SLR draws on 171 peer-reviewed publications, selected out of 2200, through a multi-stage process. The SLR intentionally has a broad scope and does not restrict itself to a particular standard or domain. This breadth of scope enables us to provide a more general and thorough analysis of the state of the art. As part of our work, we classify into a hierarchical taxonomy the various notions of evidence that we gleaned from the literature. The taxonomy, which includes 49 evidence types, is the most comprehensive classification of safety evidence built to date. This taxonomy acts as a good starting point to understand and further elaborate the evidence requirements for specific standards and specific systems, both for research and for industry.

Results from the SLR suggest that safety assurance and certification research must aim to be more rigorous from an empirical standpoint and more oriented towards industry needs. The results further provide a holistic picture of the challenges faced in relation to safety evidence, which is helpful for shaping the future research agenda on the subject.

The remainder of the paper is organized as follows. Section II presents the background for the SLR. Section III presents related work. Section IV describes the research method used. Section V presents the SLR results, and Section VI discusses the results. Section VII concludes the paper with a summary and presents future directions

II. BACKGROUND

A safety-critical system is one whose failure may cause death or injury to people, harm to the environment, or economical loss [7]. Such systems are usually subject to a rigorous safety assurance process, most commonly safety certification. The purpose of certification is to provide confidence that a system is safe for use in a specific environment under specific conditions [11]. Certification can refer to certifying the *product*, *process*, or *personnel*. Certification of product and process are usually the most challenging for software-intensive systems [31].

Confidence in safety is often achieved by establishing and satisfying safety objectives that mitigate the potential safety risks that a system can pose during operation. These objectives are typically based on one or more *safety standards* applicable to the domain(s) in which the system will operate.

Demonstrating compliance to a standard involves gathering convincing *evidence* during the lifecycle of the system to support the safety objectives defined by the standard. In general, evidence can be defined as *"The available body of facts or information indicating whether a belief or proposition is true or valid"* (Oxford Dictionary).

For a realistically large system, one can seldom argue that the evidence serves as a definitive proof that the safety objectives are met, but only that the evidence is sufficient for building (adequate) confidence in the satisfaction of the objectives. Hence, we define evidence for safety as *"information or artefacts that contribute to developing confidence in the safe operation of a system"*.

A common point of discussion in the literature concerns the nature of evidence. Some (e.g.,[28]) discuss process-based evidence (i.e., evidence about the process followed) and others (e.g., [46]) discuss product-based evidence (i.e., evidence about system characteristics). In general, the conclusion is that both types of evidence are necessary as they provide complementary perspectives of how well the product is specified and designed, and whether good process and practices went into its development.

Safety objectives and evidence are often linked by *safety arguments*, arguing that the evidence is adequate to conclude with acceptable confidence that the objectives are met. The objectives, arguments, and evidence collectively form a *safety case* [37].

III. RELATED WORK

Even though related SLRs exist in the literature (e.g., on testing [1], on requirements specification [39], and on reliability [52]), none have addressed the problem of providing evidence for safety certification. There is also prior work that studies specific types of evidence (e.g., formal methods and testing results [18]). These are complementary to the work we report here, as our aim is to develop a more general view on safety evidence.

Earlier strands of work have provided classifications of artefacts that can be used as evidence (e.g., [19]), and classifications of evidence for specific domains and standards, e.g., the nuclear domain [27] and the IEC61508 standard [43]. These have been a useful start for our work. However, they do not particularly address the topic that we aim to address with regards to evidence classification. Specifically, none of the above are targeted at developing a *unified* classification of evidence based on a systematic examination of the literature. This has led to two main gaps: first, the term evidence has largely remained a *vague* notion due to the lack of a general classification; and second, there has been little opportunity for cross-comparison of evidence requirements in different domains, standards, and systems due to the absence of a higher-level conceptual framework. The evidence classification we provide aims to address these gaps.

Similarly, previous research has reviewed techniques for evidence structuring and assessment; but the reviews are partial and not aimed at providing a comprehensive view of the spectrum of techniques that exist for these purposes. For example, there are publications that review argumentation-based assessment techniques (e.g., [23]), but these do not consider any other assessment techniques.

With regard to the challenges in the provision of evidence, most existing publications, if taken individually, motivate only the challenges that they tackle. Those that consider the challenges from a broader point of view (e.g., [25]) deal with challenges for future work and do not analyse previously studied challenges in detail. In contrast, our analysis takes a complete retrospective look at the challenges addressed and develops a more thorough picture on the research thrusts in the literature.

IV. RESEARCH METHOD

A SLR is a means of identifying, evaluating and interpreting available research relevant to a particular research question or topic area [29]. The design of the SLR reported in this paper started in October 2011. After several refinements and improvements, publication search was performed in January 2012.

The following subsections present the research questions, the data sources, search strategies, the publication selection, and the quality criteria of the SLR.

A. *Research Questions*

We formulated the following research questions:

RQ1) What constitutes evidence for safety?

This question aims to glean from the literature information and artefacts considered as evidence for system safety. The results obtained are used to develop a general classification of safety evidence types.

RQ2) What techniques are used for structuring safety evidence?

The aim of this question is to identify the structuring techniques proposed in the literature for presenting and managing safety evidence.

RQ3) What techniques are used for assessing safety evidence?

The aim of this question is to identify the techniques proposed in the literature to assess whether a collected body of evidence provides adequate confidence for a system to be deemed safe for operation.

RQ4) What challenges and needs have been the target of investigation in relation to safety evidence?

The aim of this question is to identify the perceived needs and hurdles faced in the construction, structuring, and assessment of safety evidence. In addition to providing an overall view of the topics tackled in the literature, this research question helps in identifying promising thrusts and emerging trends for future research.

B. Data Sources and Search Strategies

The search strategy included automatic search in the following publishers: ACM Digital Library, IEEE Xplore, SpringerLink, Elsevier, and Wiley.

We used the following search string to search within keywords, title, abstract and full text:

- *("critical software" OR "critical system" OR "critical systems" OR "critical equipment" OR "critical application" OR "critical applications" OR "embedded system" OR "embedded systems" OR "embedded software") AND*
- *("safety certification" OR "safety evaluation" OR "safety assurance" OR "safety assessment" OR "safety qualification" OR "safety analysis" OR "safety standard" OR "safety standards" OR "safety requirement" OR "safety requirements") AND*
- *(evidence OR "safety case" OR "safety argument" OR "assurance case" OR "dependability case").*

In addition to the automatic search, we performed a manual search on selected conferences and workshops as shown in Table 1.

C. Publication Selection

We searched for publications in peer-reviewed conferences, workshops and journals written in English that provided information about construction, structuring, and assessment of safety evidence, and/or the relevant perceived needs and challenges, all in the context of safety assurance and certification.

In Phase 1, we applied the search string to the electronic databases.

In Phase 2, the first author read the abstract of the retrieved publications to determine their relevance to the scope of the SLR. The selection criterion was to assess if the abstract referred to product-based or process-based information for demonstrating compliance with safety standards, or included the word evidence or some way to specify evidence (safety/assurance/dependability case). The first author also performed the manual searches. We included only those papers that were not identified in the automatic search. In the journals, we only considered volumes from 1990 onwards. This was the publication year of the oldest paper selected with automatic search and manual search of conferences and workshops.

In Phase 3, the first author reviewed the full text of the papers with help and guidance from the rest of the authors.

In Phase 4, the second author performed two reliability checks. First, he randomly checked over 10% of the studies of Phase 1 by reading the abstract. Second, he inspected all the papers excluded in Phase 3. Papers considered to be potentially relevant were reviewed. In addition, duplicates (papers with at least one author in common that provided equivalent answers to the research questions; e.g., an extended version of a previous paper) were removed. Other papers were added based on expert knowledge. The final number of primary studies was 171.

TABLE I. SLR PHASES AND NUMBER OF PUBLICATIONS

Source	Phase 1	Phase 2	Phase 3	Phase 4
IEEE (Publisher)	775	75	60	67
ACM (Publisher)	125	15	11	10
Elsevier (Publisher)	448	22	14	14
Springer (Publisher)	689	33	21	22
Wiley (Publisher)	163	6	4	4
Australian Workshop on Safety Critical Systems and Software	-	7	4	4
HASE (Conference)	-	0	0	0
IET System Safety (Conference)	-	12	8	8
ISoLA (Conference)	-	4	3	3
ISSRE (Conference)	-	2	2	2
SAFECOMP (Conference)	-	20	17	14
Safety Critical System Symposium (Conference)	-	14	12	12
Reliability Engineering & System Safety (Journal)	-	4	3	3
IEEE Transactions on Reliability (Journal)	-	0	0	0
IEEE Transactions on Software Engineering (Journal)	-	2	1	1
Expert knowledge	-	-	-	7
	2,200	**216**	**160**	**171**

D. Data Extraction Strategies

A data extraction template was created in a spreadsheet with respect to the research questions stated. Apart from the *bibliographic information* (title, authors, year, and publisher), we extracted from each study, *the application domain* for which safety compliance was addressed, *underlying safety standard(s)* that needed to be complied with, *information considered as evidence* in the literature, *techniques for evidence structuring*, *techniques for assessing* the evidence collected, *tool support for evidence management*, and *challenges addressed* related to evidence development, structuring, and assessment. The full information about the data extracted from the studies can be found in [38]. Each study was further evaluated across the following two quality criteria:

- **Evidence abstraction level** was assigned on the basis of the scope and specificity of evidence instances in a given study. The abstraction levels defined, from the most abstract to the most specific, were: generic, domain level, safety standard level, system type level, and (specific) system level. Using the evidence types from our evidence classification (discussed in Section V.A), example instances of evidence for the (non-generic) abstraction levels are: *Hazard specification* instantiated by [32] for the nuclear domain (domain level), *Source code* instantiated by [58] for RTCA DO178B (safety standard level), *System Historical Service Data Specification* instantiated by [56] for COTS-based systems (system type level), and *Model Checking Results* instantiated by [26] for pacemaker software (specific system level). We considered lower abstraction levels (more specific evidence) to be more useful.
- **Validation method** was assigned based on how a given study had been validated. The studies were classified as: case study (validated in real projects by practitioners different to the authors), field study (validated with data from real projects, but not during the execution of the project), action research (validated in real projects by the authors themselves), survey (validated on the basis of practitioners' opinion and perspectives), or none if no validation was given. We considered information gathered from validated work to be more useful as they better reflect the state of practice.

V. Result

This section presents the SLR results, organized according to the research questions in Section IV.A.

A. Evidence Taxonomy (RQ1)

Figure 1(a) shows the complete evidence taxonomy developed in response to RQ1, based on data extracted from the literature. Several iterations were made before the current structure of the taxonomy was developed. In each iteration, domain experts in systems safety and certification reviewed and provided feedback on the extracted evidence types. For testing results (denoted by the *Testing Results* node in the taxonomy), a suitable classification already existed in [24] and was reused.

Each leaf node in the taxonomy has been referred to by at least two of the 171 selected papers (see Section IV.C). The taxonomy is supported by a glossary, which provides a definition, a level of abstraction for each leaf node, citations to all papers in which a given leaf node appears, and a frequency ratio based on the number of these papers. Due to space constraints, we do not provide the full glossary here. The glossary and the full citations can be found in [38]. To facilitate understanding of the evidence taxonomy in Figure 1(a), we provide in Figure 1(b) definitions for a selected set of the leaf nodes, taken from the glossary.

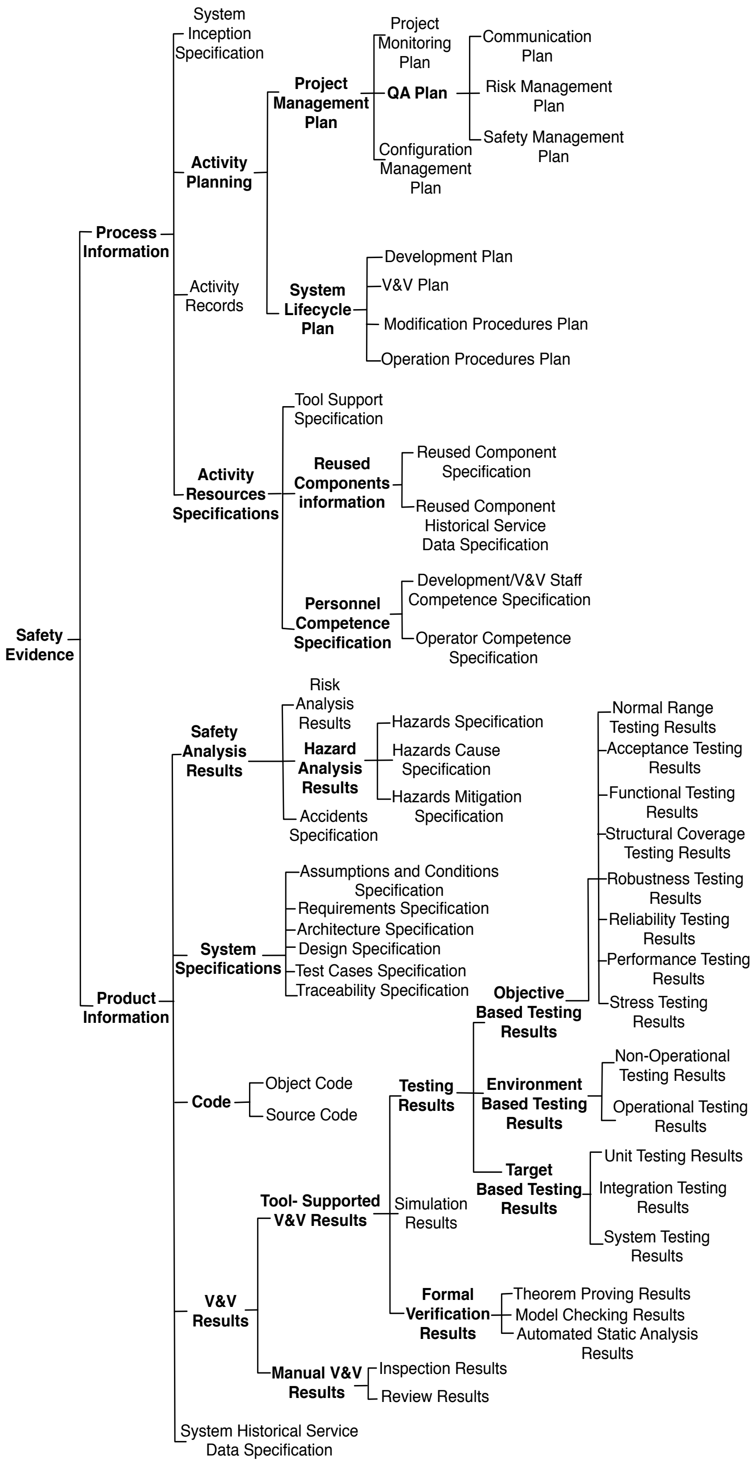
Our analysis indicates that the most frequent evidence types referred to in the literature are: *Hazards Cause Specification* (appearing in 88 out of 171 papers i.e. 52%), *Risk Analysis Results* (52%), *Hazard Specification* (43%), *Accident Specification* (34%), *Design Specification* (21%), *Requirements Specification* (21%), and *Hazards Mitigation Specification* (20%). The least frequent types are: *Communication Plan* (1%), *System Testing Results* (2%), *Object Code* (2%), *Non-operational Testing Results* (2%), and *Normal Range Testing Results* (2%). Only *Communication Plan* has not been mentioned in studies that have been validated (see Section IV.D).

The above frequencies indicate that the evidence types under *Safety Analysis Results* are the most common. These types encompass the general explanation and relationships between accidents (aka mishaps), risks, and hazards (e.g., see [11]). Several techniques and thus several finer-grained evidence types can be defined under *Safety Analysis Results*, e.g., Fault Tree Analysis for *Hazards Cause Specification* and *Risk Analysis Results* [11]. A detailed classification of safety analysis techniques and their corresponding evidence types is outside the scope of this SLR.

B. Evidence Structuring (RQ2)

Out of the 171 selected papers, 91 used or developed some technique for evidence structuring. We divide these techniques into three categories, as described below. The percentage given for each category is the ratio of papers in that category over the 91 relevant papers.

1. *Argumentation-Induced Evidence Structure (85%):* Argumentation is an approach that communicates an argument to demonstrate that a system is acceptably safe. The structure of the argumentation induces a specific structure on the evidence, as arguments need to be supported by evidence that *directly* substantiates them. In this category, GSN [54] and CAE [5], which promote a three-tiered approach composed of claims, arguments and evidence, are the most widely mentioned. The other techniques identified are Bayesian Belief Networks [6], KAOS [48], Trust Cases [2] and Safety Specification Graphs [50].
2. *Model-Based Evidence Specification (10%):* These are techniques that characterize the structure of safety evidence using models. In this category, we identified several approaches: *(1)* UML meta-modeling [43] and UML profiles built based on meta-models for specific standards such as DO-178B [58] and IEC61508 [41]; *(2)* Data modeling using entity-relationship diagrams to structure the data content in large safety cases including the evidence aspects [33]; and *(3)* Process models capturing the activities that produce and structure the evidence artifacts [21]
3. *Textual templates (5%):* These provide predefined sections or tables along with constraints for evidence structuring. One of the better-known templates for safety evidence specification is the CENELEC template [20] used in the railway domain. Other templates identified are the ACRuDA templates developed in the ACRuDA project [30] and the Template Add-ons [4],



Safety Evidence
Process Information
System Inception Specification
Activity Planning
Project Management Plan
Project Monitoring Plan
QA Plan
Communication Plan
Risk Management Plan
Safety Management Plan
Configuration Management Plan
System Lifecycle Plan
Development Plan
V&V Plan
Modification Procedures Plan
Operation Procedures Plan
Activity Records
Activity Resources Specifications
Tool Support Specification
Reused Components information
Reused Component Specification
Reused Component Historical Service Data Specification
Personnel Competence Specification
Development/V&V Staff Competence Specification
Operator Competence Specification
Product Information
Safety Analysis Results
Risk Analysis Results
Hazard Analysis Results
Hazards Specification
Hazards Cause Specification
Hazards Mitigation Specification
Accidents Specification
System Specifications
Assumptions and Conditions Specification
Requirements Specification
Architecture Specification
Design Specification
Test Cases Specification
Traceability Specification
Code
Object Code
Source Code
V&V Results
Tool- Supported V&V Results
Testing Results
Objective Based Testing Results
Normal Range Testing Results
Acceptance Testing Results
Functional Testing Results
Structural Coverage Testing Results
Robustness Testing Results
Reliability Testing Results
Performance Testing Results
Stress Testing Results
Environment Based Testing Results
Non-Operational Testing Results
Operational Testing Results
Target Based Testing Results
Unit Testing Results
Integration Testing Results
System Testing Results
Simulation Results
Formal Verification Results
Theorem Proving Results
Model Checking Results
Automated Static Analysis Results
Manual V&V Results
Inspection Results
Review Results
System Historical Service Data Specification

(a)

Accidents Specification - Specification of the events that result in an outcome culminating in death, injury, damage, harm, and/or loss as a consequence of the occurrence of a hazard of a critical system.

Activity Records - Specification of the worked performed to execute the activity planning of a critical system.

Architecture Specification - Description of the fundamental organisation of a critical system, embodied in its components, their relationships to each other, and to the environment, and the principles guiding its design and evolution.

Assumptions and Conditions Specification - Description of the constraints on the working environment of a critical system for which it was designed.

Communication Plan - Description of the activities targeted at creating project-wide awareness and involvement in the development of a critical system.

Configuration Management Plan - Description of how identification, change control, status accounting, audit, and interface of a critical system will be governed.

Design Specification - Specification of the components, interfaces, and other internal characteristics of a critical system or component.

Development Plan - Description of how a critical system will be built. It includes information about the requirements, design and implementation (coding and/or integration) phases.

Hazards Cause Specification - Specification of the factors that create the hazards of a critical system.

Hazards Mitigation Specification - Specification of how to reduce hazard likelihood and hazard consequences when a hazard cannot be eliminated in a critical system.

Hazards Specification - Specification of the conditions in a critical system that can become a unique, potential accident.

Modification Procedures Plan - Description of the instructions as to what to do when performing a modification in a critical system in order to make corrections, enhancements or adaptations to the validated system, ensuring that the required safety is sustained.

Operation Procedures Plan - Description of the instructions and manuals necessary to ensure that safety of a critical system is maintained during its use.

Project Monitoring Plan - Description of how, on a regular basis and during project execution, data about the actual progress of the activity planning of a critical system is collected and compared with the baseline plans.

Requirements Specification - Specification of the external conditions and capabilities that a critical system must meet and possess, respectively, in order to (1) allow a user to solve a problem or achieve an objective, or (2) satisfy a contract, standard, or other formally imposed documents.

Reused Component Specification - Specification of the characteristics of an existing system that is (re-)used to make up a critical system.

Reused Component Historical Service Data Specification - Specification of the dependability of a component reused in a critical system based on past observation of the behaviour.

Risk Analysis Results - Specification of the expected amount of danger when an identified hazard will be activated and thus become an accident in a critical system.

Risk Management Plan - Description of the activity regarding the development and documentation of an organised and comprehensive strategy for identifying project risks. It includes establishing methods for mitigating risk and for tracking risk.

Safety Management Plan - Description of the coordinated, comprehensive set of processes designed to direct and control resources to optimally manage the safety of an operational aspect of an organization.

System Historical Service Data Specification - Specification of the dependability of a system based on past (prior-certification) observation of the behaviour.

System Inception Specification - Specification of initial details about the characteristics of a critical system and how it will be created.

Tool Support Specification - Specification of the different tools that will be used in the system lifecycle plan.

Traceability Specification - Specification of the relationship between two or more pieces of information related to the development (process information or product information) of a critical system

V&V Plan - Description of how and by whom the V&V activities for a critical system will be executed.

(b)

Figure 1. (a) Evidence Taxonomy (b) Partial glossary for evidence types. Full glossary and citations can be found in [38] which provide a predefined set of document templates.

An important remark about the Model-Based Specification category above is that, in this category, we only consider techniques that are aimed at specifying the structure of the evidence, as opposed to the structure of the system that the evidence is for. For example, the Architecture Analysis & Design Language (AADL) [14] has been used for modeling the architecture and design of safety-critical systems, but not for modeling the structure of the systems' safety evidence. Hence, AADL was not considered. In contrast, UML, due to its broader expressive power, has been used for modeling both systems and safety evidence, and was hence considered.

C. Evidence Assessment (RQ3)

Out of the 171 selected papers, 88 used or developed some technique for evidence assessment. We classify the identified assessment techniques into four categories. The percentage given for each category is the ratio of papers in that category over the 88 relevant papers.

1. *Qualitative assessment (64%):* These are techniques that use non-numerical methods for assessment. Argumentation [54] is the most widely identified technique in this category and can be done using unrestricted natural language, (semi-) structured natural language, or graphical argumentation structures like GSN. Graphical argumentation structures have the advantage of being easier to understand, review and navigate. Argumentation can be enhanced by qualitative tags that capture the level of trustworthiness of evidence. Examples of such tags include: (1) the Safety Evidence Assurance Levels (SEALs) [15] providing four levels, the highest being incontrovertible and the lowest being supportive, to capture the degree of confidence in evidence; and (2) Safety Assurance Levels (SALs) [56] which are similar to SEALs but have the additional flexibility of allowing propagation (via propagation rules) between arguments and sub-arguments. Our review also identified qualitative methods for assessment that are not based on argumentation, e.g., the activity-based quality model of [56] which uses quality matrices to assess evidence concreteness for compliance with the IEC62304 standard. Another technique falling in this group is the evidence-confidence conversion process [57], in which safety evidence is assessed through a review process and converted into confidence on the safety on the system.
2. *Checklists (19%):* Checklists usually consist of a set of questions to be answered while reviewing evidence or a set of conditions that must be met [15]. The checklists we identified were based on the design of the system [36], based on Goal/Question/Metric approach [21], and checklists mixed with argumentation [35].
3. *Quantitative assessment (10%):* We classify techniques that use numerical measures for assessment of evidence as quantitative. Bayesian Belief Networks (BBNs) [6] are the most common in this category. Quantitative assessment can also be combined with formal argumentation structures. For example, we identified work on quantitative reasoning over safety cases using probabilities [48].
4. *Logic-based assessment (7%):* These techniques use logical formulae, such as first-order logic statements, to articulate and verify the properties of interest over evidence items and their relationships. Logic-based techniques are best suited for checking the well-formedness and consistency of evidence information. For example, the Object Constraint Language (OCL) [40] can be used to ensure that there is a consistent link between the evidence items produced for a particular system and the evidence items required by a safety standard [42].

We note that *expert judgment* can be and has been used in conjunction with all the techniques outlined above, but expert judgment per se should not be viewed as an assessment technique. This is because, for expert judgment to have any credibility, the rationale behind it must always be made explicit, e.g., through assumptions or argumentation.

D. Challenges and Needs (RQ4)

From the reviewed literature, we identified a number of general challenges and needs related to safety evidence that were common to several papers. These challenges and needs were sorted based on how many papers referred to them and are described below. Some papers noted more than one need or challenge. Full citations are found in [38].

- *Specification of evidence content:* The problem that was noted most (53 papers out of 171) was determining in a systematic way *what* information was necessary to be provided as evidence in a given domain and for a particular set of applicable standards (e.g., in [22]).
- *Construction of safety cases:* The second most identified problem (46 papers) relates to the development of safety cases, particularly providing methodological guidance for safety case construction and ways to decompose the arguments and the evidence in a way that permits more precise and cost-effective demonstration of compliance (e.g., in [5]).
- *Capturing the degree of credibility or relevance of the evidence*: We identified 26 papers in which researchers acknowledged that different evidence items could have different levels of credibility depending on their source, or different degrees of contribution towards the satisfaction of different compliance requirements (e.g., [6]). To capture credibility or relevance, one needs to be able to assign weights to the evidence items or to the links between the evidence items and the safety arguments.
- *Better development processes and better evidence about process compliance:* 21 papers noted the need

for better development processes for safety-critical systems which make it easier to rigorously verify that the development process followed is in compliance with safety standards (e.g., [21]).

- *Certification of systems made up of components and subsystems:* We identified 14 papers that mentioned challenges related to construction, structuring and assessment of evidence for systems that reuse existing components and subsystems (e.g. COTS software [56]).
- *Ambiguities in safety standards:* We identified 14 papers citing ambiguities in the standards and the existence of multiple interpretations of the evidence requirements in the standards as a source of certification issues (e.g., [12]).
- *Need for providing argumentation*: We identified seven papers that addressed the importance of demonstrating and justifying how evidence fulfills the safety requirements by argumentation (e.g., [34]).
- *Demonstration of compliance for novel technologies:* Six papers cited problems related to provision of evidence for and certification of systems that make use of technologies that are novel for safety-critical systems, e.g., adaptive systems ([51]).

E. Quality Assessment

With regards to the abstraction levels (Section IV.D), we identified that most of the studies went beyond just providing generic examples of evidence. As seen in Figure 2, the most frequent evidence abstraction level is "generic", closely followed by "safety standard level". Only 17% of the studies provide examples of evidence from/for real systems (system type level and system specific level).

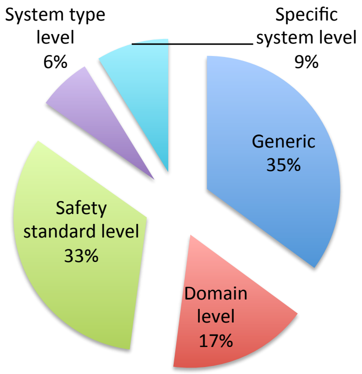


Figure 2. Percentage of studies for each evidence abstraction level

With regards to validation method (Section IV.D), the vast majority of studies (70%) have not been validated in actual projects, with practitioners, or with data from real projects. Figure 3 shows that only 13% of the studies have been validated in actual projects with action research and case studies.

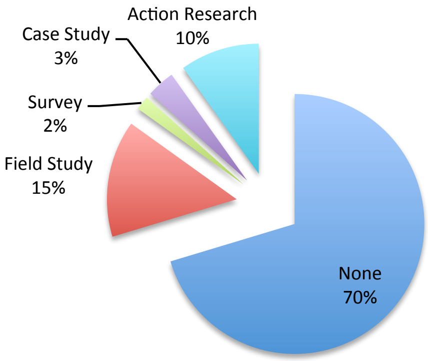


Figure 3. Percentage of studies for each validation method

VI. Discussion

This section discusses the implications of the SLR for future research and practice, as well as threats to the validity of the SLR.

A. Implications for Future Research and Practice

The review we performed provides a general view of the literature on evidence construction, structuring, and assessment along with the challenges faced in the process. The evidence taxonomy that resulted from the review depicts a holistic picture of the information and artefacts that constitute safety evidence. This taxonomy serves as a useful reference to help new researchers in the field get acquainted with the area quickly, and further can be used as basis for future research on safety evidence management tools, as such tools need to support the construction, storage, and manipulation of all the evidence types.

For the practitioners, the taxonomy helps by providing a clearer understanding of what information may be relevant for demonstration of compliance to safety standards. Particularly, information about the evidence types that have been already validated in real settings presents an advantage, as practitioners can benefit from the knowledge gained in previous applications of the evidence types by others. The taxonomy further provides a common terminology for communication about evidence requirements in the certification process. This can help reduce certification costs, as terminological differences are a common source of problems during certification [49], arising primarily due to the involvement of multiple experts who have different backgrounds and expertise.

Our results indicate that the evidence types having to do with safety analysis, design, and requirements have received the most attention in the academic literature. This prompts a follow-up investigation to confirm that these aspects are indeed the main challenges that practitioners face in real projects, and to identify potential gaps between academic research and industrial needs. Specifically, an open issue to investigate is the potential need for further research on the evidence types that were mentioned only in a low percentage of the studies. The outcome could be that either (1) more research is advisable to gain insights into the relevance and challenges associated with these types, or (2) the lack of research is due to practitioners not having recurring problems with these evidence types. Involvement and feedback from industry would be essential to determine which outcome corresponds to reality.

As indicated by the results in Section V.E, a large fraction of the evidence types found in the literature were generic (35%). We believe that more research on safety evidence at lower levels of abstraction (system type level and specific system level) would be necessary in order to gain a better understanding of concrete needs and to be able to develop more useful guidelines for practitioners.

The results obtained about the type of validation performed in the studies show that the majority (87%) of the research has not been validated in real projects. We view this as a strong indication of the need for more empirical research in the area to confirm the usefulness of the solutions proposed and to increase their impact on industrial practice.

With regards to evidence structuring (RQ2), the results are useful for both research and practice to promote further work on managing large collections of evidence data. The most widely identified evidence structuring technique is argumentation-induced structuring (Section V.B.1), which was validated in 11% of the studied papers. To further capitalize on argumentation-induced structuring, work is required on effective and modular ways to decompose general safety arguments into concrete fitness criteria, thus forming the basis for defining the evidence content and structuring it into coherent and cohesive blocks.

With regards to evidence assessment (RQ3), qualitative assessment was the technique that was identified most and was validated in 8% of the papers. To bring about industrial impact in this direction, further research is required to make qualitative reasoning more systematic, particularly when large argumentation structures are involved. In particular, work is needed on providing automated assistance during assessment to ensure correct execution of the assessment process and the soundness of assessment outcomes.

Also, we note that a large fraction of the studies that proposed techniques for evidence structuring and assessment were not validated (36% and 34%, respectively). Hence, similar to what was said about evidence types (RQ1), more empirical work is required to assess the effectiveness of the proposed techniques.

From the data extracted, we identified a total of 18 tools for evidence development, structuring, and assessment [38]. Out of these, only 4 were validated in real projects: Adelard's Safety Case Editor [3] (validated in 2 out of 5 related papers), Safety Argument Manager [44] (validated in 1 out of 3), DECOS Test Bench [2] (validated in 1 out of 2), and an unnamed tool based on Microsoft Excel to manage and generate arguments for safety [8]. Again, in line with the earlier observations, a closer examination of the usefulness and usability of the proposed tools in real industrial settings will be required.

Finally, with regards to the needs and challenges (RQ4), we note that, in the 22-year time window considered, the significant majority of the research (84%) was performed in the last 10 years. We believe this is an indication that safety evidence management, and more broadly, safety certification, is an emerging topic. To provide a finer-grained analysis of the trends, we show in Figure 3 the number of papers that tackled each of the identified challenges and needs, distinguishing papers published <=10 years ago from those published >10 years ago. The statistics suggest that two of these trends, namely demonstration of compliance for novel technologies and certification of systems made of components and subsystems, were tackled only in the last 10 years.

B. *Threats to Validity*

Discussion about the threats to validity is based on the issues regarding SLRs proposed in [16].

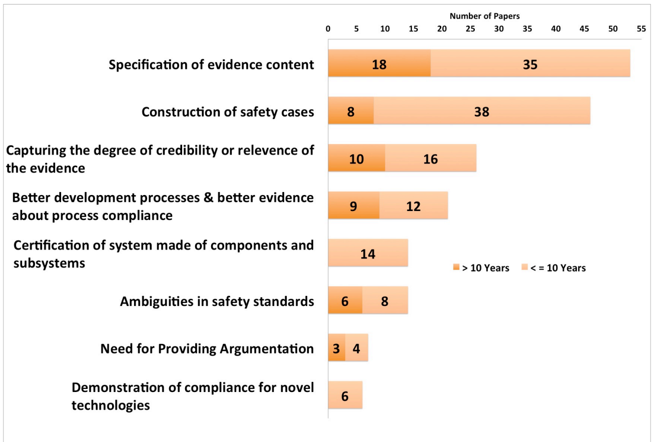


Figure 4. Identified challenges and needs

Publication bias: We began the SLR with a limited knowledge about all the related venues. Therefore, we decided to start with an automatic search. Venues and journals with the highest number of retrieved papers were then selected for manual searches.

Initially, we did not assume the breadth of the search (i.e., from 1990 until now) and considered as much peer-reviewed literature as possible. Inclusion of grey literature might have been useful as way to further ground our observations in the state of practice and thus increase internal validity. We plan to mitigate this threat in the future by validating the taxonomy, and the identified structuring and assessment techniques with practitioners.

Selection of primary studies: The first author (PhD student) performed most of the selection. This indirectly implies that, due to the lack of adequate experience, some studies might have been missed. This is a usual threat in SLRs (e.g., [16]), and reliability checks were performed to mitigate it. In addition, well-defined inclusion and exclusion criteria helped to reduce researcher bias in selection of primary studies.

Though the search string covered a wide number of studies, some studies might have been missed. Using expert knowledge for selecting studies mitigated this threat.

The criterion for publication selection (Section IV.C) helped greatly to reduce the number of studies. Although some studies were selected as a result of the reliability checks, we consider this to be logical because of a wider knowledge at Phase 4 of publication selection. The checks were performed at a final stage, after having created a first version of the evidence taxonomy and grouping all structuring and assessment techniques. Therefore, it was easier to identify evidence types, techniques and challenges.

Data extraction and misclassification: As mentioned earlier, since the first author extracted the data, we might have missed some evidence types. However, no new evidence types, techniques or challenges were found after the reliability checks, despite having a better understanding and wider knowledge.

In many cases, we had to interpret information and make assumptions about the type of information considered as safety evidence, or what validation method was being used because of the lack of details. To mitigate this threat, the first and the second authors checked, agreed upon and refined the whole set of data extracted on two occasions. We also received feedback about the taxonomy from experts.

Finally, although we might have incorrectly extracted and classified some information, we consider that having several studies supporting definition of each evidence type, techniques for structuring and assessment and challenges, mitigates this threat.

VII. Conclusions

Safety certification is an important but complex and expensive activity for most safety critical systems. A better understanding of what evidence is required for certification and how this evidence is managed and analyzed can help reduce certification costs and further make certification results more credible.

This paper presented a systematic literature review aimed at investigating the state of the art on the subject of safety evidence. In particular, the paper identified and classified the different types of evidence that have been used for demonstrating safety, and examined existing techniques for structuring and assessment of evidence. The paper further presented the range of research challenges that have been tackled in the area. The scope and quality of the studies underlying the review were analyzed and recommendations were made for addressing the gaps.

The results of the review provide useful insights for both research and practice. From a research standpoint, the challenges and gaps that have been identified are helpful for developing a future research agenda on safety evidence. Most importantly, the results suggest the need for more industry-oriented, empirical studies in the area. As for practice, the results, particularly the evidence classification developed, provide a concrete basis for learning about and tailoring the various types of evidence that practitioners need to provide in support of safety.

In the future, we would like to study how the developed evidence classification can be elaborated for different domains, to learn the strengths and weaknesses in each domain, and to apply the lessons learned for improving the certification process. We would also like to analyze the dependencies and constraints between different evidence types, and classify the evidence types according to their purpose. To further ground our evidence classification in practical needs, we plan to validate the findings of the review with the industry by means of surveys, field studies and case studies.

Acknowledgment

We acknowledge funding from the FP7 programme under grant agreement n° 289011 (OPENCOSS), from the National Research Fund, Luxembourg (FNR/P10/03 - Validation and Verification Laboratory), and from the Research Council of Norway under Project No. 203461/030. The authors would like to thank Shaukat Ali, Arnaud Gotlieb, Oleg Lisagor, Shiva Nejati, Paolo Panaroni and the industrial collaborators in OPENCOSS for their useful feedback on previous versions of the evidence taxonomy.

References

[1] W. Afzal, R. Torkar, R. Feldt, “A systematic review of search-based testing for non-functional system properties”, Info. Softw. Technol. 51(6): 957-976, 2009.

[2] E. Althammer, E. Schoitsch, Sonneck, G., Eriksson, H., Vinter, J, “Modular certification support - the DECOS concept of generic safety cases”, In: 6th INDIN, 2008.

[3] T.S. Ankrum, A.H. Kromholz, “Structured assurance cases: three common standards”, In: 9th IEEE HASE, 2005.

[4] C. Bilich, Z. Hu, “Experiences with the certification of a generic functional safety management structure according to IEC 61508”, In: Buth, B., Rabe, G., Seyfarth, T.(eds.) SAFECOMP, 2009.

[5] P. Bishop, B. Bloomfield, “A methodology for safety case development”, In: Industrial Perspectives of Safety-Critical Systems: SSS, 1998.

[6] M. Bouissou, F. Martin, A. Ourghanlian, “Assessment of a safety-critical system including software: a bayesian belief network for evidence sources”, In: Annual Reliability and Maintainability Symposium, 1999.

[7] M. Bozzano, A. Villafiorita, “Design and safety assessment of critical systems”, CRC Press, 2011.

[8] T. Dittel, H.J. Aryus, “How to “survive” a safety case according to ISO26262”, In: Schoitsch, E. (ed.) SAFECOMP, 2010.

[9] DO-178C/ED-12C, Software considerations in airborne systems and equipment certification, 2012.

[10] Draft international standard road vehicles — functional safety - ISO/DIS 26262-8, 2009.

[11] C.A. Ericson, Concise encyclopedia of system safety, Wiley (2011)

[12] J.R. Evans, T.P. Kelly, “Defence standard 00-56 issue 4 and civil standards - appropriateness and sufficiency of evidence”, In: 3rd IET International Conference on System Safety, 2008.

[13] D. Falessi,S. Nejati, M. Sabetzadeh, L. Briand, A. Messina, “Planning for safety evidence collection: a tool-supported approach based on modeling of standards compliance information”, IEEE Softw. 29(3): 64-70, 2012.

[14] P.H. Feiler, “Model-based validation of safety-critical embedded systems”, In: 2010 IEEE Aerospace Conference, 2010.

[15] J. Fenn, B. Jepson, “Putting trust into safety arguments”, In: Redmill, F., Anderson, T. (eds.) Constituents of Modern System-safety Thinking, pp 21-35, 2005.

[16] A. Fernandez, E. Insfran, S. Abrahåo, “Usability evaluation methods for the web: A systematic mapping study”. Info. Softw. Technol. 53(8): 789-817, 2011.

[17] Functional safety of electrical / electronic / programmable electronic safety-related systems (IEC 61508), 2005.

[18] A. Galloway, R.F. Paige, N.J. Tudor, R.A. Weaver, I. Toyn, J. McDermid, “Proof vs testing in the context of safety standards”, in DASC, 2005.

[19] I.M. Habli, “Model-based assurance of safety-critical product lines”, PhD thesis, University of York, 2009.

[20] I. Habli, T. Kelly, “A generic goal-based certification argument for the justification of formal analysis”. Electronic Notes in Theoretical Computer Science, 2009.

[21] I. Habli, T. Kelly, “A model-driven approach to assuring process reliability”, In: 19th ISSRE, 2008.

[22] I. Habli, T. Kelly, “Process and product certification arguments: getting the balance right”, ACM SIGBED Review 3(4): 1-8, 2006.

[23] C.M. Holloway, “Safety case notations: alternatives for the non-graphically inclined?”, In: 3rd IET International Conference on System Safety, 2008.

[24] IEEE, Guide to the Software Engineering Body of Knowledge SWEBOK (2004)

[25] D. Jackson, M. Thomas, L. Millett, Software for dependable systems: sufficient evidence? The National Academic Press, Washington D.C., 2007.

[26] E. Jee, I. Lee, O. Sokolsky, “Assurance cases in model-driven development of the pacemaker software”, In: Margaria, T., Steffen, B. (eds.) ISoLA, 2010.

[27] M. Johansson, R. Nevalainen.: “Additional requirements for process assessment in safety–critical software and systems domain”, J. Softw. Maint. Evol.: Res. Pract.. doi: 10.1002/smr.499, 2010.

[28] T. Kelly, “Can Process-Based and Product-Based Approaches to Software Safety Certification be Reconciled?” in Improvements in Systems Safety, pp 3-12. Springer, 2008.

[29] B.A. Kitchenham, S. Charters, “Guidelines for performing systematic literature reviews in software engineering” Version 2.3, EBSE Technical Report, 2007.

[30] El Koursi, E.M., Meganck, P, “Assessment criteria for safety critical computer”, In: 1998 IEEE International Conference on Systems, Man, and Cybernetics, 1998.

[31] A. Kornecki, J. Zalewski, “Certification of software for real-time safety-critical systems: state of the art”. Innov. Sys. Softw. Eng, Volume 5, Issue 2, PP 149-161, 2009.

[32] J.D. Lawrence, W.L. Persons, G.G. Preckshot, J. Gallagher, “Evaluating software for safety systems in nuclear power plants”, In: 9th Annual Conference on Safety, Reliability, Fault Tolerance, Concurrency and Real Time, Security. COMPASS, 1994.

[33] R. Lewis, “Safety case development as an information modelling problem”, In: Dale, C., Anderson, T. (eds.) Safety-Critical Systems: Problems, Process and Practice, 2009.

[34] S. Linling, T. Kelly, “Safety arguments in aircraft certification”, In: 4th IET International Conference on Systems Safety, 2009.

[35] J.A. McDermid, “Software safety: where's the evidence?”, In: 6th Australian workshop on Safety critical systems and software, 2001.

[36] D. Méry, N.K. Singh, “Trustable formal specification for software certification”, In: Margaria, T, Steffen, B. (eds.) ISoLA, 2010.

[37] Ministry of defence, defence standard 00-56 issue 4: safety management requirements for defence systems, 2007.

[38] S. Nair, J.L. Vara, M. Sabetzadeh, L. Braind, “SLR on provision of evidence for demonstrating compliance with safety standards: extracted data”, Techical Report, http://simula.no/publications/SLR_EvidenceProvision

[39] J. Nicolás, A. Toval, “On the generation of requirements specifications from software engineering models: A systematic literature review”, Info. Softw. Technol, 2009.

[40] Object Management Group (OMG). OMG Object Constraint Language , http://www.omg.org/spec/OCL/2.0/, 2006.

[41] R.K. Panesar-Walawege, M. Sabetzadeh, L. Briand, “Using UML profiles for sector-specific tailoring of safety evidence information”, ER, 2011.

[42] R.K. Panesar-Walawege, M. Sabetzadeh, L. Briand, "A model-driven engineering approach to support the verification of compliance to safety standards," ISSRE, 2011.

[43] R.K. Panesar-Walawege, M. Sabetzadeh, L. Briand, “Characterizing the chain of evidence for software safety cases: a conceptual model based on the IEC 61508 standard”, ICST, 2010.

[44] C. Pygott, S.P. Wilson, “Justifying reliability claims for a fault-detecting parallel architecture”, Journal of Systems Architecture 43(10): 735-75, 1997.

[45] Railway applications - safety related electronic systems for signalling, european committee for electrotechnical standardisation CENELEC ENV 50129, 1998.

[46] W. Ridderhof, H.G. Gross, H. Doerr, “Establishing evidence for safety cases in automotive systems - a case study”. In: Saglietti, F., Oster, N. (eds.) SAFECOMP, 2007.

[47] P. Rodriguez-Dapena, “Software safety certification: A multidomain problem”, IEEE Softw 16(4): 31-38, 1999.

[48] M. Sabetzadeh, D. Falessi, L. Briand, S. di Alesio, D. McGeorge, V. Ahjem, J. Borg, “Combining goal models, expert elicitation, and probabilistic simulation for qualification of new technology”, In: 13th HASE, 2011.

[49] M. Sabetzadeh, S. Nejati, L. Briand, and A. E. Mills, “Using SysML for modeling of safety-critical software-hardware interfaces: guidelines and industry experience”, HASE, 2011.

[50] A. Saeed, R. de Lemos, T. Anderson, “On the safety analysis of requirements specifications for safety-critical software”, ISA Transactions 34(3): 283-285, 1995.

[51] D. Schneider, M. Trapp, “A Safety engineering framework for open adaptive systems”, In: 5th International Conference on Self-Adaptive and Self-Organizing system, SASO, 2011.

[52] A. Singhal, A. Singhal, “A systematic review of software reliability studies”, Softw. Eng.: Inte. J. 1(1), 2011.

[53] M.L. Squair, “Issues in the application of software safety standards”, in Austrailian Workshop on Safety Critical Systems and Software, 2005.

[54] R. Weaver, G. Despotou, T. Kelly, J. McDermid, “Combining software evidence – arguments and assurance”, Workshop on Realising Evidence-Based Software Engineering, REBSE 2005.

[55] S.P. Wilson, T.P. Kelly, J.A. McDermid, “Safety case development: current practice, future prospect”, 12th Annul Software based system CSR workshop, 1997.

[56] F. Ye, T. Kelly, “Contract-based justification for COTS component within safety-critical applications”, In: 9th Australian workshop on Safety critical systems and software, 2004.

[57] S. Yih, C.F. Fan, “Analyzing the decision making process of certifying digital control systems of nuclear power plants”, Nuclear Engineering and Design 242: 379- 388, 2012.

[58] G. Zoughbi, Briand, L., Labiche, Y, “Modeling safety and airworthiness (RTCA DO-178B) information: conceptual model and UML profile”. Softw. Sys. Model, 2011.