SAFETY4RAILS

EVALUATION METHODOLOGY

Deliverable 8.1

Lead Author: UIC

Contributors: Fraunhofer, MdM, EGO, RFI, LDO, CEIS, IC, UREAD, ETRA, WINGS, LAU, PRO, CdM, FGC, TCDD

Dissemination level: PU - Public

Security Assessment Control: passed

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 883532.

D8.1 EVALUATION METHODOLOGY	
Deliverable number:	8.1
Version:	1.0
Delivery date:	17/12/2021
Dissemination level:	PU - Public
Nature:	Report
Main author(s)	UIC
Contributor(s)	CdM, CEIS, EGO, ETRA, FGC, Fraunhofer, IC, LAU, LDO, MdM, PRO, RFI, TCDD, UREAD, WINGS.
Internal reviewer(s)	Fraunhofer (*also for final security assessment control*), UREAD.
External reviewer(s)	European Railway Information Sharing & Analysis Center (ER-ISAC) *(in process, feedback input to future work)*

Document control			
Version	Date	Lead author(s)	Change(s)
0.1	29/07/2021	UIC	Draft structure.
0.2	02/09/2021	UIC	Initial version.
0.3	20/10/2021	UIC	First draft with contribution from UIC, CEIS.
0.4	26/11/2021	UIC	Second draft with contribution from UIC, CEIS, IC, ETRA, MDM.
0.5	07/12/2021	UIC	Final draft with contribution from FGC, Fraunhofer, UMH, LAU.
0.6	13/12/2021	UIC	Final draft ready for review.
0.7	16/12/2021	UIC	Preliminary final version after review from WINGS, UREAD, Fraunhofer.
0.8	17/12/2021	Fraunhofer	Further update with Fraunhofer additional input and formatting. Identification in this table of “Lead author”.
1.0	17/12/2021	Fraunhofer	Creation of V1.0 from V0.8.

DISCLAIMER AND COPYRIGHT

ABOUT SAFETY4RAILS

SAFETY4RAILS is the acronym for the innovation project: **Data-based analysis for SAFETY and security protection FOR detection, prevention, mitigation and response in trans-modal metro and RAILway networkS**. Railways and Metros are safe, efficient, reliable and environmentally friendly mass carriers, and they are becoming even more important means of transportation given the need to address climate change. However, being such critical infrastructures turns metro and railway operators as well as related intermodal transport operators into attractive targets for cyber and/or physical attacks. **The SAFETY4RAILS project delivers methods and systems to increase the safety and recovery of track-based inter-city railway and intra-city metro transportation.** It addresses both cyber-only attacks (such as impact from WannaCry infections), physical-only attacks (such as the Madrid commuter trains bombing in 2004) and combined cyber-physical attacks, which are important emerging scenarios given increasing IoT infrastructure integration.

SAFETY4RAILS concentrates on rush hour rail transport scenarios where many passengers are using metros and railways to commute to work or attend mass events (e.g. large multi-venue sporting events such as the Olympics). When an incident occurs during heavy usage, metro and railway operators must consider many aspects to ensure passenger safety and security, e.g. carry out a threat analysis, maintain situation awareness, establish crisis communication and response, and they have to ensure that mitigation steps are taken and communicated to travellers and other users. **SAFETY4RAILS will improve the handling of such events through a holistic approach.** It will analyse the cyber-physical resilience of metro and railway systems and deliver mitigation strategies for an efficient response, and, in order to remain secure given everchanging novel emerging risks, it will facilitate continuous adaptation of the SAFETY4RAILS solution; this will be validated by two rail transport operators and the results will support the re-design of the final prototype.

TABLE OF CONTENT

List of tables

List of figures

Executive summary

This document is the deliverable D8.1 – Evaluation Methodology– of SAFETY4RAILS, aiming at presenting the methodology which is going to be applied for the evaluation of the SAFETY4RAILS Information System (S4RIS) platform that will be tested during the four simulation exercises to be held between M14 (January 2022) and M19 (June 2022).

This evaluation methodology is based on the user perspective and mainly focuses on the impact of the S4RIS platform in enhancing resilience against combined cyber-physical threats to railway infrastructure and metro systems.

First, some existing methodologies taken from past projects and based on open-source research are reviewed. This first overview is completed by an overview of evaluation methodologies used and/or developed by the SAFETY4RAILS end-users and members of the Consortium.

Then, based on the literature review, an evaluation methodology framework is described for the SAFETY4RAILS project. This proposed methodology is based on: i) borrowing from the UK FSR guidelines on validation; and ii) also on answering these four questions: “what, who, how and when” is relevant. The evaluation will be mainly performed by the end-users of the project participating in the exercises and will focus on 2 main aspects:

- The organisation of the exercise.
- The performance of the S4RIS against pre-defined objectives related to:
    - Usability.
    - Specific requirements laid out by the end-users in SAFETY4RAILS Deliverable D1.4.
    - Scenario-based requirements/objectives to be identified in SAFETY4RAILS Deliverable D8.2, (referenced back to e.g. tool specific requirements/specifications identified in D1.4).

The requirements/objectives relevant for the evaluation of the S4RIS by the end-users have been identified among the 300 requirements described in SAFETY4RAILS Deliverable D1.4. For each category of requirements/objectives identified, a set of questions (closed-ended and open-ended) is prepared. Depending on the questions, different tools can be used, among them questionnaires, debrief and group-based techniques.

Finally, the report provides a first attempt to tailor the methodology to the first 2 exercises which are already described in SAFETY4RAILS Deliverable D8.2. The last 2 exercises will be described in a later stage within task 8.2 on “Operational simulation exercises – performance”.

The methodology described in this deliverable will served as a guide within task 8.3 on “Evaluation - End-user and developer feedback for improvement”.

1. Introduction

1.1 Overview

This deliverable presents the methodology which is going to be applied for the evaluation of the four simulation exercises that will be held between M14 (January 2022) and M19 (June 2022) to test the SAFETY4RAILS Information System (S4RIS) platform in operational environments. This evaluation methodology is based on the user perspective and mainly focuses on the impact of the S4RIS platform in enhancing resilience against combined cyber-physical threats to railway infrastructure and metro systems.

The evaluation methodology described in this document as part of WP8 is to be distinguished from the evaluation methodology within WP6 (Implementation of SAFETY4RAILS Information System), which will be performed solely by the technical partners.

The main objective of this document is to describe how the scenarios and simulation exercises will be evaluated by the end-users, ensuring the quality and consistency of the analysis and evaluation.

1.2 Structure of the deliverable

This deliverable is structured as follows:

- Section 1 introduces the deliverable.
- Section 2 provides an overview on existing evaluation methodologies from the end-user perspective.
- Section 3 presents the methodology framework that will be applied within SAFETY4RAILS.
- Section 4 describes the methodology applied to the first two simulation exercises.
- Section 5 provides the main challenges and the future work.
- Section 6 contains the bibliography.
- Section 7 regroups the annexes.

2. Review of existing evaluation methodologies

2.1 Introduction

This section aims to provide a first overview of identified relevant evaluation methodologies informed from past projects and based on open-source research. Evaluation and validation methodologies focusing on evaluating and/or validating a tool are presented at first followed by evaluation methodologies for exercises. This first overview is completed by an overview of evaluation methodologies used and/or developed by the SAFETY4RAILS end-users and members of the Consortium.

2.2 Literature review on evaluation methodologies

2.2.1 Implementation research

"Implementation research" is a term used to describe the scientific study of processes used in the implementation of initiatives and the contextual factors that affect these processes[1]. Implementation research is based on a set of pre-defined categories, which sort the attributes of the tool being studied. These are acceptability, adoption, appropriateness and feasibility and are described in more detail below:

- **Acceptability** is defined by the perception among stakeholders that the tool is agreeable, e.g. the stakeholders perceive that there is a benefit to using the tested tool over other types of tools.
- **Adoption** categorises the intention or willingness to use the tool.
- **Appropriateness** describes the perceived fit or relevance of the tool in a particular setting or for a particular target audience or issue. It is related to key words such as compatibility, trialability, suitability, usefulness, and practicability.
- **Feasibility** is related to the practicality and fit of the tool, i.e. the extent to which the tool can be used in a particular setting or organisation.

Then, within each relevant category, “success criteria” must be defined. For example, success criteria associated with acceptability could include “the tool is user friendly”, where user friendly would need to be defined as a function of the tool.

2.2.2 Design Science Research Methodology Evaluation phase

Design Science Research Methodology (DSRM) [2] enables the development, design and building of new artifacts. A key part of DSRM is the evaluation of the artefact and this consists of rigorously demonstrating that the artefact achieves its goals. While there is no one specific evaluation methodology recognised as part of DSRM, the field has several possible ways of going forward. The Framework for Evaluation in Design Science (FEDS) was developed to aid Design Science researchers with the evaluation phase of the methodology. FEDS is made up of four steps: (1) explain the goals of the evaluation, (2) choose the evaluation strategy or strategies,

---

[1] D.H. Peters, N.T. Tran, and T. Adam, *Implementation research in health: a practical guide*. World Health Organization, 2013

[2] vom Brocke J., Hevner A., Maedche A. (2020) Introduction to Design Science Research. In: vom Brocke J., Hevner A., Maedche A. (eds) Design Science Research. Cases, Cham. https://doi.org/10.1007/978-3-030-46781-4_1

(3) determine the properties to evaluate, and (4) design the individual evaluation episode(s)[3]. Before the evaluation, Pries-Heje et al. (2008) recommend that the following three questions[4] be answered:

1. What is going to be evaluated?
2. How is it going to be evaluated?
3. When is it going to be evaluated?

2.2.3 STEP (Systematic Test and Evaluation Process)

Systematic Test and Evaluation Process (STEP) [5] was designed for software evaluation in order to measure the quality of the system. In order to carry out STEP, one must plan, acquire the testware (detail test objectives, test sites, etc.) and measure the behaviour (execute the tests and evaluate).

"The STEP process defines evaluations according to three main phases: (1) Scoping and Test Strategy, (2) Test Preparation, (3) Testing, Results, and Final Report, and a fourth, optional phase (4) Integration and Deployment that is determined by the sponsor on a case-by-case basis. Each STEP phase has different objectives, actions and associated document deliverables"[6]. An important aspect of the methodology is that the evaluation criteria and test strategy need to be defined before installing or testing the evaluation products. This includes the identification of a full set of evaluation criteria that the products will be tested against and the definition of the scenario tests that will be performed.

2.2.4 User-centred evaluation

The User-centred evaluation proposed by Borland (2000)7 consists of an extensive usability evaluation and trials. User-centred evaluation puts emphasis on the role of the user rather than the system and considers the needs and limitations of the end-users. The focus lies in testing the system or specific modules in a near-real-life scenario by giving test persons realistic tasks in a staged environment. According to Borland, there is no standard user-centred evaluation method; in fact, the respective methodology needs to be chosen according to the specific use case scenario and to the specific functionality that needs to be tested.

---

[3] Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: a Framework for Evaluation in Design Science Research. European Journal of Information Systems, 25, 77-89.
[4] J. Pries-Heje, R. Baskerville, and J.R. Venable, *Strategies for Design Science Research Evaluation*. In: Proceedings of the 16th European Conference on Information Systems (ECIS 2008), Galway, Ireland, 2008.
[5] Craig, R. Jaskiel, S. « Systematic Software Testing. » Artech House, 2002.
[6] Sarah Brown (2007),Standardized Technology Evaluation Process (STEP) User's Guide and Methodology for Evaluation Teams
[7] Borland, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. Journal of Documentation 56(1):71-90.

2.2.5 Usability evaluation on basis of ISO 9421-11 definitions for ergonomics of human-system interaction

One of the quantitative methods to determine usability is outlined in ISO 9241-11, a standard consisting of specific metrics about how well a user fulfils specific goals[8] (see Figure 1). This standard includes the main concepts of user-centred design[9].

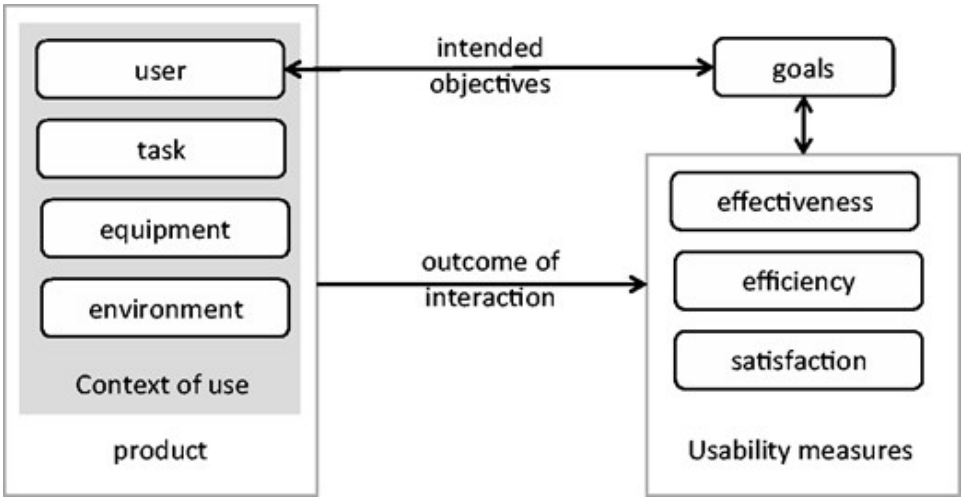


FIGURE 1 USABILITY FRAMEWORK ACCORDING TO ISO 9241–11

ISO 9241-11 describes in depth how users should interact with a product, employing hands-on methods to indicate its overall usability via three attributes, which are[10]:

- **Effectiveness**: To what extent the user can achieve a goal with accuracy and completeness.
- **Efficiency**: The level of effort and resource usage which is required by the user in order to achieve a goal in relation to accuracy and completeness.
- **Satisfaction**: The positive associations and absence of discontent that the user experiences during the performance.

If the indicated measures of effectiveness, efficiency, and satisfaction are fulfilled adequately, the product can be considered to have attained an acceptable level of usability[11]. In 2016, ISO published a revised version of standard ISO 9241-which defines usability as the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [12]. In particular, ISO defines effectiveness as the "accuracy and completeness with which users achieve specified goals," efficiency as the "resources used in relation to the results achieved" and satisfaction as the "person's perceptions and responses that result from the use of a system, product or service"[13].

---

[8] International Organization for Standardization. Ergonomic requirements for office work with visual display terminals (VDTs) Part 11 Guidance on usability. International standard, 9241-11. Geneve: ISO; 1998

[9] International Organization for Standardization. Ergonomics of human-system interaction - Part 210: human-centred design for interactive systems. International standard, 9241-210. Geneve: ISO; 2010

[10] Bevan N. International Standards for HCI. In Claude G, ed. Encyclopedia of Human Computer Interaction. Hershey, PA, USA: IGI Global; 2006: 362 – 372.

[11] International Organization for Standardization. Ergonomic requirements for office work with visual display terminals (VDTs) Part 11 Guidance on usability. International standard, 9241-11. Geneve: ISO; 1998.

[12]International Organization for Standardization Ergonomics of human-system interaction: part 11: usability: definitions and concepts (ISO/DIS 9241-11.2:2016). German and English version pr EN ISO 9241-11:2016

[13] International Organization for Standardization Ergonomics of human-system interaction: part 11: usability: definitions and concepts (ISO/DIS 9241-11.2:2016). German and English version prEN ISO 9241-11:2016

Standards such as the ISO 9241-11 are especially suitable to apply to new technologies and applications. In support of this notion, Bevan (2009) concludes that these standards should be used more frequently in usability work as they define good practice, are objective, can ensure consistency in the work, and can provide benchmarks for intervention by designers.[14]

2.2.6 Usability evaluation through questionnaires

A questionnaire is a research tool. The quality and accuracy of the characteristics surveyed depends on the questionnaire itself, the rating scale it uses, the way it is filled in, and the process of completing it.[15]

Two of the most popular free-to-use, standardised questionnaires which assess perceived usability are considered here[16]. Other types of questionnaires were not found to be suitable for the evaluation cause of either being too rudimentary e.g. (USE, UMIX-LITE) or to extensive and not free of charge (e.g. SUMI).[17]

Option A: The System Usability Scale (SUS) developed by Brooke (1996) to evaluate the usability of various technical systems. The questions consist of 10 items measured on a Likert scale to quantify the perceived usability by the participants. It uses direct and reversed items and a simple coding scheme which gives a total usability score for the assessed system[18] (See Annex II).

Option B: The Computer System Usability Questionnaire (CSUQ) in its third version is a 16-item questionnaire with its roots in the Post-Study System Usability Questionnaire (PSSUQ) developed at IBM in the 1980s[19].The CSUQ is identical to the PSSUQ, but due to the change in research context slight changes to the wording were applied. The items produce four scores – one overall and three subscales. The rules for computing them are[20]:

- Overall: average the responses for items 1–16 (all the items)
- System Usefulness (SysUse): average items 1–6
- Information Quality (InfoQual): average items 7–12
- Interface Quality (IntQual): average items 13–15

The resulting scores follow a 7-point Likert Scale plus a non-applicable (NA) option with lower scores indicating a higher degree of satisfaction[21]. The sub-scales provide a more detailed breakdown of different factors affecting the tool.

2.2.7 Usability evaluation through group-based techniques:

The use of group-based techniques has also been identified as a useful methodology for usability evaluation[22].

When research requires the integration of professional opinions and feedback to reach a consensus among a group of experts, it is common to collect the experts' viewpoints through conferencing, e.g., brainstorming, focus groups, group interaction, online meetings, etc.

---

[14] Bevan, N (2009). International standards for usability should be more widely used. Journal of Usability Studies 4(3),106–113.
[15] Zafiropoulos K. (2005): How to write down a scientific paper. Athens. Kritiki.
[16] James R. Lewis (2018): Measuring Perceived Usability: The CSUQ, SUS, and UMUX, International Journal of Human–Computer Interaction, DOI:10.1080/10447318.2017.1418805
[17] Assila, A., Oliveira, K.M., & Ezzedine, H. (2016). Standardized Usability Questionnaires: Features and Quality Focus. Computer Science and Information Technology, 6
[18] Brooke, J. (1996). SUS: A ''quick and dirty" usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, A.L. (Eds.), Usability Evaluation in Industry. Taylor and Francis, London.
[19] Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction, 7, 57–78.
[20] Lewis, J.R. (2018). Measuring Perceived Usability: The CSUQ, SUS, and UMUX. International Journal of Human–Computer Interaction, 34, 1148 - 1156.
[21] Sauro, J., & Lewis, J. R. (2016). Quantifying the user experience: Practical statistics for user research (2nd ed.). Cambridge, MA: Morgan Kaufmann.
[22] Chai, C. S., & Der-Thanq, V. C. (2004). A review on usability evaluation methods for instructional multimedia: an analytical framework. *International Journal of Instructional Media*, *31*(3), 229.

These methodologies allow individual subjects to participate together as a group and for the researcher to collect data. The interactions that take place through group-based techniques add depth to the information obtained and enable shared meanings to emerge, which is not possible in individual interviews. Group techniques also tend to stimulate the expression of new, creative and unbiased ideas. The most commonly used group methodology is the Focus Group. Despite the advantages to group-based techniques mentioned above, they have some limitations.

The most frequently mentioned disadvantages are associated with group dynamics. Group discussions may be dominated by one or several individuals or influenced by the researcher's expectations. Participants may feel pressured to conform to their peers, to a sub-group majority, to the dominant people or to the ideas expressed by someone with authority in the group. Group dynamics may influence the attitudes of individual participants. Groups can impede individual reactions, resulting in "groupthink" and sometimes in "group polarisation" (with two totally opposite views diverging). It is difficult to ensure equal participation among all group members. In addition, the central role of the researcher can fundamentally influence a group discussion.

2.2.7.1 Nominal group technique (NGT)

The Nominal Group Technique (NGT) was developed to respond to these limitations. NGT is a highly structured group-based technique using face-to-face meetings. NGT combines individual and group phases. The purpose of the structure and individual phases is to limit group dynamics and social power dynamics. The technique prevents dominant individuals from controlling the group and limits the researcher's interaction in the generation of ideas.

The NGT is a single-purpose technique and can only cover a limited number of topics and issues. The question posed at the beginning of the NGT meeting is critical. It will determine the quality of the ideas generated.

The NGT follows a format consisting of the following phases: the silent generation of ideas in writing; feedback from group members takes the form of voting on ideas, which provides quantitative results of a qualitatively generated idea. This voting makes it possible to identify the ideas that have the highest importance for the group, because they have achieved the highest scores, as well as to identify those ideas that have been selected by the majority of the participants even if they have not given the highest scores. The result is therefore very transparent and objective, all participants are taken into account in a proportionate way and the researcher's influence on the analysis of the results is minimal. Given the qualities mentioned above, this technique is well suited for usability evaluation, and its application has already been demonstrated in this field[23].

2.2.7.2 Delphi method

The Delphi method also addresses some of the limitations mentioned about focus groups. It is based on the idea that a consensus could be achieved among a panel of experts through multiple iterations of a questionnaire with controlled feedback (Dalkey and Helmer 1963). This methodology is based on open-ended questionnaires. Then, based on participants' responses, the questionnaires are modified or developed in a series of rounds to seek greater consensus. In this series of rounds, the opinion of the participants may change as the perspective of the other participants is included in the development of the new questionnaires. This methodology and its adaptations have also proven useful in usability evaluation[24].

---

[23] Vertesi, A., Dogan, H., & Stefanidis, A. (2020). Usability Evaluation of Virtual Learning Environments: A University Case Study. In *Online Teaching and Learning in Higher Education* (pp. 161-183). Springer, Cham.

[24] Hsin-Ke, L., Sung-Chun, T., Peng-Chun, L., Kuo-Chung, C., & Chen, A. N. (2020). Toward a new real-time approach for group consensus: A usability analysis of synchronous Delphi system. *Group Decision and Negotiation*, *29*(2), 345-370; CAN, Gülin Feryal; DEMIROK, Seda. Universal usability evaluation by using an integrated fuzzy multi criteria decision making approach. *International Journal of Intelligent Computing and Cybernetics*, 2019. Dawood, K. A., Sharif, K. Y., Ghani, A. A., Zulzalil, H., Zaidan, A. A., & Zaidan, B. B. (2021). Towards a unified criteria model for usability evaluation in the context of open-source software based on a fuzzy Delphi method. *Information and Software Technology*, *130*, 106453.

2.2.8 Validation guidance provided by the UK's Forensic Science Regulator

In the deliverable D1.1 Project Management Manual (hereafter referred to as "D1.1") validation guidance provided by the UK's Forensic Science Regulator (hereafter referred to as "UK FSR") was introduced.[25] "Validation involves demonstrating that a method used for any form of analysis is fit for the specific purpose intended, i.e. the results can be relied on".[26] The UK FSR's guidance is focussed on validation of technical methods and procedures used by forensic units.[27] SAFETY4RAILS is not focussed on forensics, but some of SAFETY4RAILS data later in operational systems (e.g. after the project at TRL9) has the potential to be the basis for expert evidence in criminal justice systems[28] and in addition, in the rail and metro domain end-users need also to be able to rely on the results. SAFETY4RAILS evaluation and validation strategy can be informed by the UK FSR validation guidance. The Fraunhofer Society, represented by its institute for Applied Optics and Precision Engineering (IOF) and the SAFETY4RAILS project coordinator specifically have applied this validation guidance successfully in an earlier H2020 innovation project.[29]

As presented in D1.1, Figure 2 "provides a simplified representation of how end-users requirements can be perceived to be at the centre of both the first initiation of research for the development of a method (cycle bottom left), the actual development of a method (bottom right) and the validation of a method (top). In this representation: the increase in TRLs for a new method would be in the same order i.e. lowest TRLs bottom left and highest TRLs top; and validation of a final method (top) is separated from the development of a method.

[25] SAFETY4RAILS, D1.1 Project Management Manual, V1.0, October 2020, p. 9.
[26] UK Forensic Science Regulator, Guidance: Validation (FSR, Issue 2, 2020), p.5.
[27] UK Forensic Science Regulator, Codes of Practice and Conduct For Forensic Science Providers and Practitioners in the Criminal Justice System (FSR, Issue 7, 2021) p. 47.
[28] Supra: D1.1, p.10.
[29] Crabbe S., Lucas M., Ramm R., Sgrenzaroli M., Developmental Validation of the 3D-Forensics system, Version 1.1, September 2020.

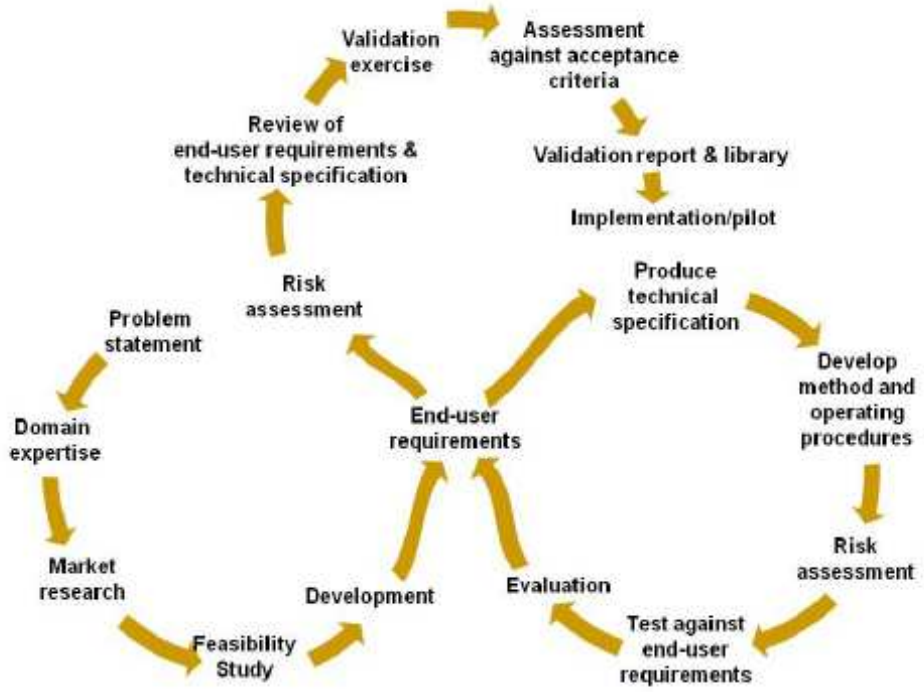


FIGURE 2 END-USER REQUIREMENTS – IMPORTANCE AND PLACE IN DISTINCT CYCLES[30]

In criminal justice systems, it is critical that validation remains the responsibility of end users and that it is separate from the method development, at least for those issues which could influence the results. (Manufacturers can however support users with their own "developmental validation" of the final version of a method.)"[31] Developmental validation is the acquisition of objective evidence of the fitness of purpose for a new or novel methodology often performed by the developer or manufacturer.[32] Developmental validation encompasses larger more in-depth studies.[33]

The UK FSR validation framework is represented in FIGURE 3.

[30] Image from UK Forensic Science Regulator, *Guidance: Validation* (FSR, Issue 1, 2014), p.19.
[31] Supra: D1.1 page 9.
[32] UK Forensic Science Regulator, Guidance: Validation (FSR, Issue 2, 2020), p.17.
[33] Op.cit.

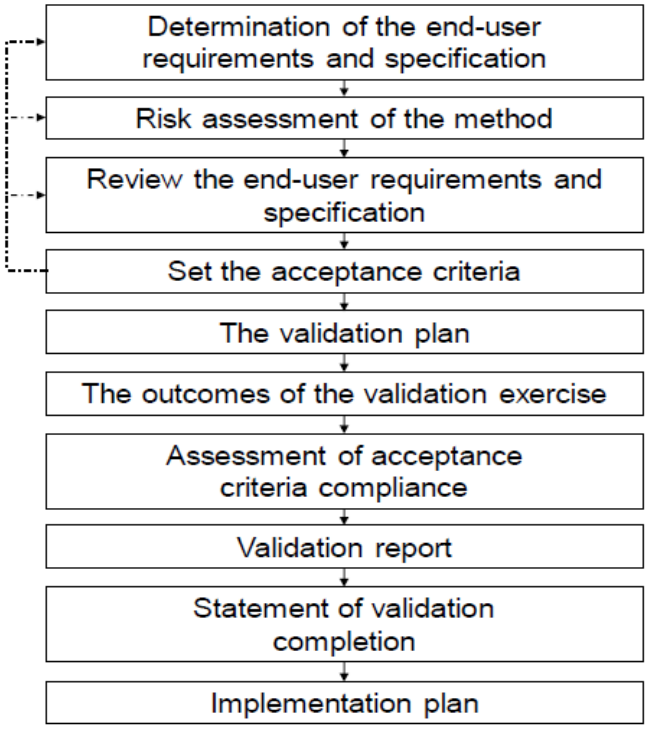


FIGURE 3 VALIDATION PROCESS FRAMEWORK PUBLISHED IN UK FSR CODES[34]

The UK FSR however accepts that "When a method has been validated in another organization the forensic unit shall review validation records to ensure that the validation performed was fit for purpose. It is then possible for the forensic unit to only undertake verification for the method to demonstrate that the unit is competent to perform the test/examination." (ILAC).[35]

"The Regulator defines verification thus: "Confirmation, through the assessment of existing objective evidence or through experiment, that a method, process or device is fit (or remains fit) for the specific purpose intended. There is an overriding requirement that there is evidence that the provider's own competent staff can perform the method at the given location."[36]

In essence, taking the last points together, forensic units need to use validated technical methods and procedures. When a technical method and/or procedure has been validated by another forensic unit, further forensic units need to be able to verify their ability to use the method and/or process correctly. The validation of an initial technical method and/or procedure can be supported by evidence provided through "developmental validation" performed by a developer or manufacturer. Validation of a forensic method and/or procedure indicates availability for operational use which would be equivalent to TRL9 - actual system proven in operational environment.[37]

The UK FSR also provides more detailed specific guidance on method validation for digital forensics. Here under the sections dealing with the scale of validation needed for novel methods, the guidance is given:

"A novel method using new software tools will include the sort of validation and verification procedures dictated in software engineering to demonstrate that the software development was to the required standard. Appropriate standards ensure that the software's internal engineering is correct. Therefore, there should be

---

[34] *Image from UK Forensic Science Regulator, Method Validation in Digital Forensics (FSR, Issue 2, 2020), p.11.
[35] Supra: UK FSR, Validation Issue 2 p.6 with reference to the International Laboratory Accreditation Cooperation (ILAC), Modules in a Forensic Science Process, paragraph 3.10, p.16
[36] Op. cit.
[37] European Commission, HORIZON 2020 – WORK PROGRAMME 2014-2015 General Annexes, G. Technology readiness levels (TRL).

evidence of use of a formal development method and/or a quality management system, as well as evidence of unit and system testing, including test plans and results."[38]

Also, with specific reference to the quote from ILAC in the (4th paragraph above), the guidance is given:

"The above description is often referred to as verification; in reality it is performance verification with a key proviso that the validation records have been reviewed first. To review the existing validation records implies that:

a. There is something to review the validation records against (i.e. an end-user requirement and technical specification);

b. There is access to the validation records in sufficient detail to assess against the end-user requirement, specification and risk assessment; and

c. The method is the same or demonstrably comparable."[39]

Applying the above to SAFETY4RAILS:

- SAFETY4RAILS will not be able to implement a full validation process for the S4RIS and contributory tools following the guidelines from the UK FSR as:
  - SAFETY4RAILS target is TRL7; and
  - (We evaluate that we would not have the resources to carry out the comprehensive validation process based on the UK FSR guidelines for the S4RIS platform and all contributory tools).

- SAFETY4RAILS can "borrow" from the UK FSR approach to validation
  - SAFETY4RAILS has been doing this e.g. initial steps in the project were to set end-user requirements and specifications and to review end-user requirements and specifications from contributory tools as also presented in the D1.4 Specification of the overall technical architecture, section 2.1.[40]
  - The WP6 and specifically T6.4 is focussed on gathering evidence for "developmental validation" i.e. objective evidence of the fitness of purpose for the S4RIS platform and its contributory tools performed by the developers. It will entail larger more in-depth studies of the core technical requirements and specifications. Here validation and verification procedures dictated in software engineering to demonstrate that the software development was to the required standard are expected to be documented (if not already done in earlier deliverables) as well as unit and system testing, including documenting test plans and results (again, if not already done in earlier deliverables). (*Task 6.4 was started earlier in month 13 and its detailed activities are in planning*)
  - In WP8 the simulation exercises will identify in advance which requirements/specifications to be tested in each simulation exercise and the end-users acceptance of the demonstrated results will be evaluated. In principle, the "acceptance criteria" will be the meeting of the relevant specification(s). End-users will also be requested / have the opportunity to propose revisions and/or additions to the requirements and specifications defined to date. These revisions and/or updates may still be considered during the project, depending on resource limitations. They will also be input into steps after the project to implement the results such as validation of products following also the UK FSR guidelines even more comprehensively.

[38] UK Forensic Science Regulator, Method Validation in Digital Forensics, (FSR, Issue 2, 2020), p.23.
[39] Op. cit. p.24.
[40] SAFETY4RAILS, D1.4 Specification of the overall technical architecture, V1.0, October 2021, p. 14.

- SAFETY4RAILS evaluation and validation results from WP6 and WP8 can contribute to further validation in later TRL steps.

2.3 Literature on exercise evaluations

Based on an overview of existing literature[41] related to the organisation and the evaluation of exercises (including both in simulated environments and on the field), three key phases and their main steps appear common to prepare the exercise evaluation are described below:

1. Ahead the exercise
    - Determination of exercise objectives, that can be conducted following the SMART criteria (e.g., Specific, Measurable, Achievable, Realistic and Task-related).
    - Definition of the evaluation team, that should include one leader/coordinator in addition to several evaluators.
    - Mobilisation or development of the necessary resources (training, briefings, materials) to evaluators to perform their assignment.
    - Drafting of a key questions list for each determined exercise objectives and of assessment and measure means - both could be included into one document as an evaluation plan.
2. During the exercise
    - Ensure the coverage of evaluations aspects by evaluators, based on the objectives of the exercise.
    - Reporting of observations and information into evaluators documentation.
    - Dedicate a timeslot for direct feedback just after the exercise.
3. Following the exercise
    - Assessment of objectives achievements, that may rely on the so-called OAJR assessment criteria (Observation, Analysis, Judgement, Recommendations).
    - Organisation of post-exercise meetings, as an opportunity to resolve potential issues or problems that occurred during the exercise.
    - Preparation of the post exercise report, including for instance exercise outcomes (describing the ratio performance/objectives), exercise management focusing on the quality of the process, next steps foreseen.

2.4 End-user experiences

This section aims at describing the main steps that are commonly taken by the end-users for the evaluation of security solutions.

A consultation with project end-users was performed regarding existing evaluation methodologies for security solutions.

According to the feedback by the questioned project end-users, no standardised methodology for the evaluation of new information technology tools in the transport sector has been identified. Existing evaluation methodologies that have been described by CDM, FGC, PRORAIL and PKP/UIC (in annex 7.2) are mostly based on risk assessment methodologies using the evolution of the number of occurrences of security incidents and KPIs such as time for service recovery or delays caused by the incidents.

The end-users evaluated new security solutions, by applying certain varying KPIs and their expert knowledge to the object to be assessed. Most often, end-users tend to perform evaluations based on the set of criteria laid

[41] "A framework for major emergency management", Department of Housing, Local Government and Heritage, September 2021 (URL); "Tasmanian Exercise Framework Evaluation Plan Exercise Transfundo", Tasmania Fore Service, May 2013, (URL); "Emergency telecommunications table-top simulation guide", ITU Publications, 2020 (URL); "Handbook Evaluation of Exercises", Swedish Civil Contingencies Agency, March 2011 (URL); "Handbook 3 Managing Exercises", Australian Disaster Resilience Handbook Collection", 2012 (URL);

out in the call for tender. However, based on the testimonies of the end-users and their experience, it can be assumed very generally that following steps are taken to evaluate new security technology tools:

1. Current situation assessment - conduct an overview of the existing technology taking into account their strengths and weaknesses. The focus here is to comprehend which areas need higher level of security and/or which technologies currently used could be augmented or supported by other solutions.
2. Define needs and expectations – describe the requirements towards new technology very clearly. Focus on technological and functional aspects.
3. Determine the budget
4. Investigate the market and available solutions
5. Announce call for tenders (e.g., Request for Quote; Invitation to Offer) describing especially:
    a) Defined requirements – they should be based on what users try to achieve, not on a particular technical solution (that way novel solutions can be found).
    b) Defined evaluation criteria, e.g.
        - Capabilities – how the technology capability meets the needs?
        - Security – is the technology compliant with the latest security standards?
        - Is it a long-term solution?
        - Flexibility.
        - Integration capabilities – to be able to integrate a program/system with the one currently used.
        - Cultural alignment – to determine how to work with a technology partner in the long-term.
        - Cost of services – to be aware of the maintenance, setup and potential licensing costs
        - Technology acceptance
    c) The specification and any appropriate document to the specification.
    d) Test phase period – to evaluate selected technologies in the test environment.
    e) The level of detail used for requirements: either draw up a detailed list of requirements (including their importance) or describe them in a general way and focus on the fundamental goals.
    f) Others.
6. Set the deadline for the receipt of tenders.

3. Methodology for SAFETY4RAILS

3.1 Introduction

Based on the literature review, the evaluation methodology based on i)borrowing from the UK FSR guidelines on validation; and ii) also on answering these four questions: “what, who, how and when” is relevant for the SAFETY4RAILS project. The approach under ii) is based on the Design Science Research Methodology where there are three questions but with the addition of the “who”. While the answers to which will vary depending on the given scenario studied, here in this chapter, potential approaches to answering are presented.

3.2 What is going to be evaluated?

The main output of the SAFETY4RAILS project is the SAFETY4RAILS Information System (S4RIS). S4RIS is an integrated platform that offers and combines risk assessment, monitoring, simulation and decision support capabilities as well as “visualisation means to prevent, forecast, detect, defuse, respond and mitigate the impact of cyber and physical threats in a holistic methodological and operational approach resulting in a collaboration between cyber-physical security technologies and actors”[42]. The SAFETY4RAILS project aims at a prototype of the S4RIS which can be demonstrated and validated in an operational environment. The overall philosophy is to bring different technologies together and combine these with the S4RIS, to provide various functionalities towards supporting the end-users in the railway and metro sector in the handling of cyber, physical and combined cyber-physical threats.[43]

Four simulation exercises, which represent 4 scenarios, will be organised within the project to test and evaluate the S4RIS platform. The simulation will be carried out between January 2022 (project month 16) and July 2022 (project month 22), with time between the simulations to implement identified potential for improvement of the developed information system.

For each scenario, the tool capabilities that can be provided either through the S4RIS platform or as a standalone (in the first exercises) will be described in D8.2[44] (first version – development of a blueprint exercise handbook).

The main objective of the evaluation is to identify concerns, strengths and areas for improvement.

The evaluation will focus on 2 main aspects:

- The organisation of the exercise (as carried out).
- The performance of the S4RIS against pre-defined objectives related to:
    - Usability.
    - Specific requirements laid out by the end-users in SAFETY4RAILS Deliverable D1.4.
    - Scenario-based requirements/objectives to be identified in SAFETY4RAILS Deliverable D8.2.

[42] SAFETY4RAILS Grant Agreement, version 1.0

[43] SAFETY4RAILS Deliverable D1.4

[44] SAFETY4RAILS Deliverable D8.2

As identified in section 2.2.8, the evaluation is expected to include here an assessment/opinion on how far the requirements/specifications tested were met and if necessary any proposals for revisions and/or additions to the requirements and specifications defined to date.

3.2.1 Evaluation of the organisation of the exercise

The feedback of all the participants will be helpful for preparing the next exercises as well as future simulation exercises. The main focus will be on what can be done differently for the next exercises or what improvements need to be made.

3.2.2 End-users Requirements

Within the first period of the project, over 300 end-user requirements have been identified as the basis for the development of the S4RIS platform considering the resilience of metro and rail infrastructure with the Smart City concept of multi-modality broadly.

All these requirements are documented and a specification in answer to each requirement is provided in SAFETY4RAILS Deliverable D1.4[45]. As stated in D1.4,"*The requirements and specifications are input into both the S4RIS development cycle in SAFETY4RAILS and also future evaluation and validation cycles. The requirements and specifications have been formulated for a future S4RIS product."*

Beyond the fact that D1.4 serves as a reference for the main development and integration of tools and components for the S4RIS, it is also the basis for the technical evaluation and validation of the S4RIS and the evaluation of the S4RIS in operational environments during simulation exercises.

The distinction between the technical/developmental evaluation and validation and the evaluation of the S4RIS in operational environments during simulation exercises is carried out in close consultation with WP6 (task 6.4) to identify and evaluate the affiliated requirements inherent within the work packages.

The identification of the requirements for the evaluation of the S4RIS in operational environments during simulation exercises is based on:

- The priority level defined in D1.4:
  - ***"Essential** - This implies that a future product will not be acceptable unless these requirements are provided in an agreed manner.*
  - ***Conditional** - This implies that these are requirements that would enhance the product but would not make the product unacceptable if they were absent.*
  - ***Optional** - This implies a class of functions that may or may not be worthwhile."*

- The feasibility/relevance of the requirement in the context of the use-case based exercises and the set of data that can be analysed.

From a general perspective, three kinds of requirements relevant for the evaluation of the S4RIS during simulation exercises have been identified: Usability, S4RIS platform specifics, and scenario-based requirements/ KPIs (referenced back to e.g. tool specific requirements/specifications identified in D1.4)**.**

3.2.2.1 Usability

One objective in this task is the **evaluation of the usability** as part of the user experience. As stated in the review of existing methodologies ISO 9241-11 (for ergonomic of human-system interaction) defines usability

[45] SAFETY4RAILS Deliverable D1.4

as the “extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use”[46].

As part of the usability, the Graphical User Interface (GUI) will be evaluated. Most of the requirements identified in D1.4 are technical and will be validated within WP6 in the technical validation. The evaluation from the end-users performed within WP8 will focus on the ease of use (for all technical requirements), relevance (GUI-R06, GUI-R07, GUI-R17, GUI-R23) and the overall end-user satisfaction.

TABLE 1 USABILITY RELATED REQUIREMENTS FROM D1.4

Req.-ID	Short name	Key objectives	Type of user requirement	Priority rank
GUI-R01	**Web-based interface**	to enable easy access to S4RIS	**Technical**	Essential
GUI-R02	**Login page**	to define log-in window	**Technical**	Essential
GUI-R03	**Single point of access to the tools**	to have all tools available for use in a single page	**Technical**	Essential
GUI-R04	**Grouping of tools**	to group tools based on their area of use	**Technical**	Essential
GUI-R05	**How to launch tools**	to define how each tool will be accessed by the operator	**Technical**	Essential
GUI-R06	**Display of tools based on user role**	to guarantee that only authorised users can launch the tools	**Technical**	Essential
GUI-R07	**Tools keywords and short descriptions**	to help users to easily understand what a tool is used for	**Technical**	Essential
GUI-R08	**Log-out button**	to defined log-out position	**Technical**	Essential
GUI-R09	**Home page button**	to define position of home page button	**Technical**	Conditional
GUI-R10	**Account management**	to enable the user to manage their account and change their own password	**Technical**	Essential
GUI-R11	**Settings and configuration**	to enable editing of setting and configuration	**Technical**	Essential
GUI-R12	**Language**	to enable changing of the displayed language	**Technical**	Essential
GUI-R13	**Bar with additional functions**	to quickly and easily find additional S4RIS functions and menus.	**Technical**	Essential
GUI-R14	**Opening web-based tools**	to define how to open tools with web-based interface	**Technical**	Essential
GUI-R15	**Opening desktop tools**	to define how to open tools with desktop application	**Technical**	Essential
GUI-R16	**Opening CLI tools**	to define how to open tools with Command Line Interface only	**Technical**	Conditional

[46] International Organisation for Standardisation Ergonomics of human-system interaction: part 11: usability: definitions and concepts (ISO/DIS 9241-11.2:2016).

Req.-ID	Short name	Key objectives	Type of user requirement	Priority rank
GUI-R16a	**Opening CLI tools - BB3d**	to define how to deal with BB3d	**Technical**	Conditional
GUI-R16b	**Opening CLI tools**	CaESAR - to define how to deal with CaESAR	**Technical**	Conditional
GUI-R16c	**Opening CLI tools**	SARA - to define how to deal with SARA	**Technical**	Conditional
GUI-R17	**User confirmation on certain actions**	to let the user correct some unwanted actions	**Technical**	Essential
GUI-R18	**Font type and size**	to ensure readability	**Technical**	Conditional
GUI-R19	**Error display**	to inform the user of errors	**Technical**	Essential
GUI-R20	**S4RIS account creation**	to enable operators to request the creation of an account	**Technical**	Conditional
GUI-R21	**Help and documentation**	to provide access to provide tutorials and/or documentation	**Technical**	Conditional
GUI-R22	**Frequently/recently used tools**	to collect the tools frequently/ recently used	**Technical**	Conditional
GUI-R23	**Dashboard**	to display information to the user	**Technical**	Conditional
GUI-R24	**Mobile interface**	to enable usage of S4RIS from mobile devices	**Technical**	Conditional

3.2.2.2 S4RIS platform specifics

Apart from evaluating the usability and aside from the tool specific requirements, the S4RIS platform specifications, as described in SAFETY4RAILS Deliverable D1.4, represent a commitment to the **performance of the items to be developed** and this will be evaluated (see SAFETY4RAILS D1.4: Req-ID P-01 – P-17 -).

Twenty requirements have been described in SAFETY4RAILS Deliverable D1.4. Most of them will be validated primarily during the technical/developmental validation phase (WP6). Six requirements, described below, will be particularly considered in the evaluation by the end-users during the exercises under WP8.

The evaluation will focus on the integration features of the S4RIS platform concerning:

- Data input (P02).
- Integration of the tools' capabilities and outputs (P04-P05).
- Integration in the User environment (P03).

The online manual and the skill needed will also be assessed.

TABLE 2 S4RIS PLATFORM SPECIFIC REQUIREMENTS FROM D1.4

Req.-ID	*Short name*	*Key objectives*	*Type of user requirement*	*Priority rank*
P-02	**Consolidation of end-user inputs**	The S4RIS platform shall provide a solution whereby similar input which is required by S4RIS contributory tools is only inputted by the user once and this input is then shared between tools needing it.	**Functional**	Conditional
P-03	**End User configuration**	The S4RIS platform shall provide a solution whereby the end-user can configure the platform for his/her specific infrastructure.	**Functional**	Essential

Req.-ID	Short name	Key objectives	Type of user requirement	Priority rank
P-04	**Minimum requirements for S4RIS use**	The S4RIS platform shall provide the end-user directions as to which input and in what granularity is needed from them as preconditions in order to be able to use the S4RIS and each contributory tool (and their combinations)	**Functional**	Essential
P-05	**Identification of useful S4RIS contributory tool combinations**	The S4RIS platform shall provide an indication to the user of which combinations of tools (and in which order) have the potential to provide promising results to them, including the likely scope of the results.	**Functional**	Essential
P-15	**Manual**	The S4RIS platform provide an on-line manual / help function	**Functional**	Essential
P-16	**Skill / training**	The S4RIS platform shall be designed for operators with classic education of railway operators and/or crisis managers and with a training of 2 days / hours. Individual tools shall be designed for operators with additional training of not more than 1 day per tool.	**Functional**	Essential

3.2.2.3 Scenario-based requirements (referenced back to e.g. tool specific requirements/specifications identified in D1.4)

During each simulation exercise, an integrated version of the S4RIS platform will be evaluated along with the capabilities that will be simulated in the context of the scenario used during the simulation exercise.

In SAFETY4RAILS D8.2 “First version – Development of a blueprint exercise handbook exercise handbook”, the simulation exercises are described in detail. It includes the description of the tool capabilities (i.e. specifications in answer to requirements) that will be tested for each resilience stage of the scenario with the specific objectives of the simulation and the expected performance to be evaluated.

The main objective of this evaluation, especially during the exercises, is to provide feedback to the solution providers on the possible improvement of the tools.

3.3 Who is going to evaluate?

The evaluators are the potential end-users of the S4RIS platform, who are representatives from transport operators and infrastructure managers (especially rail and metro). Local authorities and police representatives who are the main responsible for managing the crisis in case of a terrorist attack may also be also involved in the evaluation.

The end-users can be divided into 3 categories depending on their involvement in the project and in the exercise:

- End-user representatives organising the exercise (and therefore actively using the tool):
  - A detailed evaluation of the results with several iterations of the simulation will be organised per tool/capability.
- End-user representatives from the Consortium (and therefore observing the pilot case):
  - The evaluation will focus on the usability of the S4RIS platform and the outputs provided.
- End-user representatives outside the Consortium and mainly from the Advisory Board (and therefore observing the pilot case):
  - The evaluation will focus on the simulation exercise in general and the results achieved.

3.4 How is it going to be evaluated?

3.4.1 Observations (open or structured)/Physical quantitative measures

During the exercise the participants will be asked to perform a systematic observation. What each end-user representative not organising the exercise will need to observe will be laid out ahead of the exercise. This could take the form of, for example, observing how long it took for a tool to be launched ready for operation, say, the Tool X was launched within 2 minutes. The defined objectives will act as the criteria for whether or not the associated requirements were met.

3.4.2 Questionnaires

For each category of requirements identified in section 3.2, a questionnaire, to be filled-in by the end-users, will be prepared. In each questionnaire, the name, the company and the position of the responders will be requested. The respondents will also be asked if they are willing to participate in a follow-up interview. The questionnaire will use both open ended and closed questions. For closed questions, a Likert scale is to be used. To minimise the risk to the integrity of the unbiased answers given in the questionnaire, it will take place as soon as the exercises are finished and before a debrief is conducted.

In the context of SAFETY4RAILS project, through the use of a standardised and established questionnaire, the results accomplished during the four exercises can be compared with each other. The questionnaire conducted will aim at quantifying the perceived usability of the SAFETY4RAILS Tool by the end-user.

3.4.2.1 Questionnaire for the evaluation of the organisation of the exercise:

Who: all participants.

When: immediately after the simulation.

Evaluation criteria examples:

- Closed-ended questions to be answered between 1 (strongly disagree) to 5 (strongly agree):
    1. The exercise was well-structured and organised.
    2. The process was clearly explained.
    3. The exercise format and the lengths of the planning time was appropriate.
    4. The debrief session at the end of the simulation was useful.
- Open-ended questions:
    5. Which part of the exercise did you find the most useful and why?
    6. Were there any parties missing whose participation would have given added value for the exercise?
    7. What can be done differently for the next exercises or what improvement need to be made?
    8. What are the main lessons learnt for you and why?

3.4.2.2 Questionnaire for the evaluation of the usability of S4RIS GUI:

Who: all end-users.

When: immediately after the simulation.

Evaluation criteria examples:

- Closed-ended questions to be answered between 1 (strongly disagree) to 5 (strongly agree):
    1. S4RIS platform is easy to use.
    2. It's easy to find the information needed.

3. It's easy to understand what a tool is used for.
4. It is always clear for me what I should do.
5. Overall, I'm satisfied with the system demonstrated for this scenario.

- Open-ended questions:
  6. What could be improved in the GUI
  7. What complexity / functions are not necessary and can be deleted or reduced?
  8. What could be improved to make the handling more transparent?
  9. Any proposals for revisions and/or additions to the requirements and specifications defined to date?

3.4.2.3 Questionnaire for the evaluation of S4RIS platform specific:

Who: all end-users.

When: immediately after the simulation exercise.

Evaluation criteria examples:

- Closed-ended questions to be answered between 1 (strongly disagree) to 5 (strongly agree):
  1. The functions in S4RIS are well integrated.
  2. The S4RIS platform helps the user to choose the right combinations of tools for managing the situation.
  3. The combinations of tools simulated provide relevant results.
  4. The S4RIS platform can be adapted for my specific infrastructure and scenarios.
  5. The S4RIS platform provide an on-line manual / help function which is easy to understand.
  6. The results from the combination of tools does not pose risks to the company's internal policies (e.g. on privacy).
  7. The results from the combination of tools does not pose ethical or psychological risks to the user.
  8. The system presented would provide added value for my organisation especially for addressing combined cyber-physical threats
- Open-ended questions:
  9. Were there situations where you did not understand what the system was doing?
  10. Would you recommend the system presented to your colleagues and why?
  11. What could be improved in the context of this scenario?
  12. Any proposals for revisions and/or additions to the requirements and specifications defined to date?

3.4.2.4 Questionnaire for the evaluation of the scenario-based requirements:

Who: end-users: experts from the end-user company organising the exercise and end-users from the Consortium attending the exercise.

When: after the simulation for each resilience stage.

Evaluation criteria examples:

- S4RIS tools Capabilities:

  - **Prevention phase**
    - Description of the objectives of each tool capability in the context of the scenario

- Questions for each objective:
  - Closed-ended questions to be answered between 1 (strongly disagree) to 5 (strongly agree):
    1. The objective was successfully met.
    2. The output will help for the prevention phase.
    3. The GUI of the individual tools is user-friendly.
  - Open-ended questions:
    4. Which aspects in the GUIs presented were too complex and not helpful?
    5. What is the added value to the prevention phase that you know from your current daily work?
    6. What could be improved in the context of this scenario?
    7. Any proposals for revisions and/or additions to the requirements and specifications defined to date?

- **Detection phase**
  - Description of the objectives in the context of the scenario
  - Questions for each objective:
    - Closed-ended questions to be answered between 1 (strongly disagree) to 5 (strongly agree):
      1. The objective was successfully met.
      2. The time for processing was acceptable.
      3. The GUI of the individual tools is user-friendly.
    - Open-ended questions:
      4. Which aspects in the GUIs presented were too complex and not helpful?
      5. What would be your acceptable time to be processed?
      6. What is the added value to the detection phase that you know from your current daily work?
      7. What could be improved in the context of this scenario?
      8. Any proposals for revisions and/or additions to the requirements and specifications defined to date?

- **Response phase**
  - Description of the objectives in the context of the scenario
  - Questions for each objective:
    - Closed-ended questions to be answered between 1 (strongly disagree) to 5 (strongly agree):
      1. The objective was successfully met.
      2. The time for processing was acceptable.
      3. The outputs will help for the decision-making process.
      4. The GUI of the individual tools is user-friendly.
    - Open-ended questions:
      5. Which aspects in the GUIs presented were too complex and not helpful?
      6. What would be your acceptable time to be processed?
      7. What is the added value to the response phase that you know from your current daily work?
      8. What could be improved in the context of this scenario?
      9. Any proposals for revisions and/or additions to the requirements and specifications defined to date?

- **Recovery phase**
  - Description of the objectives in the context of the scenario
  - Questions for each objective:
    - Closed-ended questions to be answered between 1 (strongly disagree) to 5 (strongly agree):

1. The objective was successfully met.
2. The output will help for recovering quickly.
3. The GUI of the individual tools is user-friendly.

- Open-ended questions:
  4. Which aspects in the GUIs presented were too complex and not helpful?
  5. What is the added value to the recovery phase that you know from your current daily work?
  6. What could be improved in the context of this scenario?
  7. Any proposals for revisions and/or additions to the requirements and specifications defined to date?

- **Overall**
  - Description of the overall objectives in the context of the scenario
  - Questions:
    - Closed-ended questions to be answered between 1 (strongly disagree) to 5 (strongly agree):
      1. The combination of tools contribute to improving the resilience of the system.
    - Open-ended questions:
      2. Which capabilities are the most important/useful for this scenario?
      3. What are the current obstacles for adopting such a system?
      4. What could be improved in the context of this scenario?
      5. Has any limitation of tools been discovered during the exercise? If so, please specify.
      6. What is the overall added value as may be assessed from your own experience in your current daily work?
      7. What were the main lessons learnt by you and why?
      8. Any proposals for revisions and/or additions to the requirements and specifications defined to date?

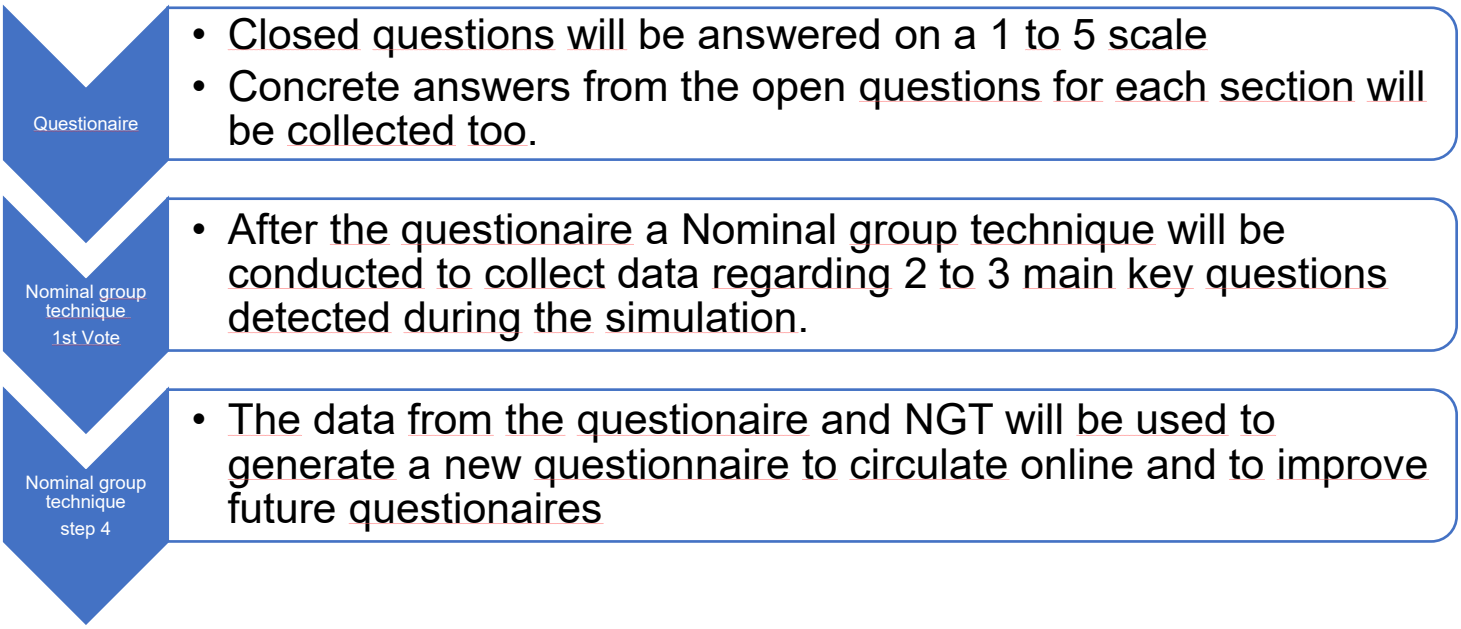
3.4.3 Debrief

Right after each simulation, a debrief will be carried out to collect feedback from the participants on strengths and areas for improvement.

3.4.4 Group-Based techniques

For the participants who agree, a Nominal Group Technique (NGT) will be organised to collect data for the most relevant open questions (2 to 3 questions). The data collected though questionnaires and the NGT will be analysed and processed for a follow-up questionnaire.

3.4.5 Data collection Process resume



3.5 When is it going to be evaluated?

According to the description of action, the evaluation will be conducted as part of four simulation exercises which will take place in Madrid, Ankara, Rome and Milan. During the exercises, the determined objectives and their associated requirements will be assessed by the users and observers. Followed by the exercises, the participants will be provided with the questionnaire on the evaluation of the exercises, the usability of the S4RIS platform and the S4RIS platform specific requirements. A debrief will be then organised for the open-ended questions.

The Scenario-based requirements will be evaluated after the simulation of each resilience stage. The closed-ended questions will be raised through an online questionnaire (for example using slido) whereas the open-ended questions will be addressed through Nominal Group Techniques.

4. Evaluation methodology applied to SAFETY4RAILS exercises

This section represents a first attempt to tailor the methodology to the first 2 exercises which are already described in SAFETY4RAILS Deliverable D8.2. The last 2 exercises will be described in a later stage within Task 8.2.

It is likely that during a given exercise, the full set of all parameters (including the full set of specifications in answer to requirements) will not be measurable. This is due to the particularities of the scenarios and that not all tools in S4RIS are applicable to all simulation exercises. The Evaluation Framework should serve as a guideline to be considered for setting up the exercises.

4.1 Exercise 1 (Madrid, Spain)

4.1.1 Objective of the exercise

The objective of the simulation exercise that will be held in Madrid is to evaluate both the first version of the S4RIS platform in the context of a Cyber-physical attack during a football game and the individual tool capacities. The scenario is described in detail in D8.2.

The simulation exercises will involve several S4RIS capabilities to cover each resilience stage in the context of scenario: prevention, detection, response, recovery.

In this simulation exercise, 11 tools will be deployed to provide some of their functionality . Some functions will be integrated into the S4RIS platform whereas others will be stand-alone. Full integration in the S4RIS platform is planned to be completed by the next simulation exercises.

4.1.2 What is going to be evaluated?

This section lays out the specific requirements that will be evaluated in this scenario. As laid out in Chapter 3, the organisation of the exercise will also be evaluated and all evaluation processes laid out therewithin carried out.

4.1.2.1 Requirements for the Prevention phase

The functional requirements identified in SAFETY4RAILS Deliverable D1.4 and the corresponding objectives in the scenario that have been identified in SAFETY4RAILS Deliverable D8.2 are the following:

TABLE 3 MDM SIMULATION EXERCISE - REQUIREMENTS FOR THE PREVENTION PHASE

No	***Req.-ID - from D1.4***	**Short name**	**MDM Scenario objectives**	**Inte-grated in S4RIS (Y/N)**
MDM-PRE-1	BB3d_01	Bomb blast loading	Provide bomb blast simulations in order to understand how a bomb could affect the metro infrastructure, particularly the tunnels and the development of an event. This information will further support the Civil Construction Department in MDM for building more resilient physical structures (e.g. the tunnels) and reduce damage to passengers.	N

No	*Req.-ID - from D1.4*	Short name	MDM Scenario objectives	Inte-grated in S4RIS (Y/N)
MDM-PRE-2	CaESAR_02	CaESAR should identify weak points in the railway/metro system	The weakest/most critical components and associated cascading effects will be identified. An overall resilience analysis of the infrastructure will be done before the event	N
MDM-PRE-3	CaESAR_05	Implementation and evaluation of mitigation measures	Assist the end-user in selecting the best mitigation measures to respond against different possible events in the MDM infrastructure	N
MDM-PRE-4	CAMS_02	Maintenance and repair budget calculation	The individual tool will be used to inform the metro operator to allocate to repair/maintain/ rehabilitate the infrastructure after a set of possible events, therefore providing the necessary input to make a proactive plan and be ready in case of an attack. The metro operator will also be provided with information regarding the asset condition and degradation due to normal ageing, enabling timely response ahead of malfunctioning.	Y
MDM-PRE-5	DATA FAN-2	High prediction performance of results, e.g. anomaly detection	Provide information about the expected number of passengers to happen on the day of the football match. The end-user will be able to run what-if scenarios to analyse how they will affect the number of passengers and delays in the infrastructure (e.g. the closure of a station).	N
MDM-PRE-6	iCrowd_02	Simulate an evacuation because of terrorism (bomb, gas release) or natural disaster (fire/flood)	Provide simulation capabilities to understand better the chances of detection during infiltration/escape per configuration (camera and guards locations) and infiltration/escape total times.	*Y*
MDM-PRE-7	iCrowd_04	Detect blind-spots because of guards' movements and insufficient cameras	Revealing blind spots and other related vulnerabilities in case of a threat actor trying to escape	Y
MDM-PRE-8	PRIGM_04	PRIGM should give service for end nodes and create outputs for end-users	Provide detailed report regarding vulnerabilities and attack surfaces within the system (mainly hardware-based attacks), supporting Network Security Expert or Cybersecurity Officer in the definition and development of countermeasures against cyber and/or cyber-physical attacks.	N
MDM-PRE-9	RAM2_01	RAM2 should provide risk assessment and prioritisation	Provide vulnerability and security gaps assessment, along with risk assessment for each of the operational units in the metro system.	Y

No	Req.-ID - from D1.4	Short name	MDM Scenario objectives	Inte-grated in S4RIS (Y/N)
MDM-PRE-10	SECURAIL_3	Computation of Risk	Enable off-line risk analysis of the metro infrastructure to understand the level of risk for each critical asset during a given hazardous event	Y
MDM-PRE-11	TISAIL_2	Detection of cyber-threats related to the railway sector: Internet-Exposed Assets and credential leakage	Provide an assessment of the vulnerabilities and exposed assets of the infrastructure related to Workstations, PCs, CCTV systems and Power Grid	Y

4.1.2.2 Requirements for the Detection phase

The functional requirements identified in SAFETY4RAILS Deliverable D1.4 and the corresponding objectives in the scenario that have been identified in SAFETY4RAILS Deliverable D8.2 are the following:

TABLE 4 MDM SIMULATION EXERCISE - REQUIREMENTS FOR THE DETECTION PHASE

No	Req.-ID - Need.-ID	Short name	Objective for the MDM exercise	Inte-grated (Y/N)
MDM-DET-1	CuriX_02	Catalogue-Based Outage Prevention	Crisis Manager will be alerted when deviations from normal behaviour (anomalies) or potentially upcoming disruptions of technical systems (IT and OT) from their monitoring data are detected. The crisis manager can check metrics and which technical devices are responsible for causing the major change in the system behaviour.	Y
MDM-DET-2	CuriX_03	Infrastructure Monitoring (including cyber threats)	The crisis manager can monitor the health of the monitored technical system.	Y
MDM-DET-3	DATA FAN-7	Manner of the applied anomaly detection	Data gathered regarding the flow of passengers will be used to detect significantly high passenger volumes in stations and trains, also considering days with really crowded events	N
MDM-DET-4	RAM2_02	RAM2 should generate correlated insights	Correlation of data gathered from multiple monitoring sources in order to detect potential threats. For example, it will be able to correlate the different attack vectors happening in the station	Y
MDM-DET-5	TISAIL_4	Detection of cyber-threats related to the railway sector: Vulnerabilities	Inform the Crisis Manager about possible threat actors targeting CCTV	Y
MDM-DET-6	TISAIL_5	Detection of cyber-threats related to the railway sector: Spear Phishing	Inform the Crisis Manager about possible spear-phishing campaigns targeting mail domains of the MDM personnel.	Y

No	Req.-ID - Need.-ID	Short name	Objective for the MDM exercise	Inte-grated (Y/N)
MDM-DET-7	WINGS_03	Support of A.I. techniques	Analyse anomalies in the train speed so that an alert can be sent to the system team/driver. Check if there is an overcrowded area in the facility and raise an alert.	N

4.1.2.3 Requirements for the response phase

The functional requirements identified in D1.4 and the corresponding objectives in the scenario that have been identified in D8.2 are the following:

TABLE 5 MDM SIMULATION EXERCISE - REQUIREMENTS FOR THE RESPONSE PHASE

No	Req.-ID - Need.-ID	Short name	Objective for the MDM exercise	Inte-grated (Y/N)
MDM-RES-1	CaESAR_05	Implementation and evaluation of mitigation measures	Evaluate mitigation steps regarding their influence on the resilience, including cascading effects computation. As a pre-condition, CAESAR will count with the system topology provided by SecuRail.	N
MDM-RES-2	DATA FAN-2	High prediction performance of results, e.g. anomaly detection	Predict the passenger load in real-time in other stations once another is closed, helping to better respond the situation.	N
MDM-RES-3	iCrowd_01	Simulate realistic crowd congestion levels	Crowd simulator providing advanced insights regarding crowd movement and behaviour for a set of boundary conditions related to the event.	Y
MDM-RES-4	RAM2_01	RAM2 should provide risk assessment and prioritisation	Risk-based prioritisation of issues, case management for tracking response actions. End user consumes the data through RAM2 Dashboards display. The user follows the prioritised alerts and mitigation steps for each of the alerts for risk reduction and response to detection of ongoing threats.	Y
MDM-RES-5	WINGS_03	Support of A.I. techniques	Provide details, alerts of the detected issue in the train speed to aid the response action. Alerts are also raised in the case of overcrowded areas and guidelines in case of evacuation are provided.	Y

4.1.2.4 Requirements for the recovery phase

The functional requirements identified in SAFETY4RAILS Deliverable D1.4 and the corresponding objectives in the scenario that have been identified in SAFETY4RAILS Deliverable D8.2 are the following:

TABLE 6 MDM SIMULATION EXERCISE - REQUIREMENTS FOR THE RECOVERY PHASE

No	*Req.-ID - Need.-ID*	Short name	Objective for the MDM exercise	Inte-grated (Y/N)
MDM-REC-1	BB3d_01	Bomb blast loading	Safety managers in the metro system will leverage the information provided by the bomb blast simulations in order to create mitigation countermeasures (e.g. safety distance, protective hardening, etc.). Number of casualties and people injured for out-door bomb attack scenarios are provided.	N
MDM-REC-2	CAMS_10	Assessment of recovery	Crisis Manager will be provided with time and cost needed to respond to the crisis and restore normal functioning, so that resource deployment and reaction is based on proactive actions planned. Railway operator will be aware of vulnerability and fragility of the asset after the incident, so to improve resource deployment and control financial loss in the future.	Y
MDM-REC-3	RAM2_03	RAM2 should provide alert and insight mitigation steps	Provide mitigation steps for each alert raised by the system.	N

4.1.3 Who is going to evaluate?

Overall, registration for the exercise is ongoing and the following is simply a preliminary list of evaluators' companies:

- All participants to the exercise
- End-user representatives organising the exercise:
  - The Metro of Madrid security team and other relevant departments.
- End-user representatives from the Consortium:
  - Rail partners: FGC, PRORAIL, RFI, TCDD.
  - Metro partners: EGO.
  - Local authority: CDM.
- End-users representatives outside the Consortium and mainly from the Advisory Board: a dedicated workshop will be organised for them.

4.1.4 How is it going to be evaluated?

The evaluation will be conducted through observations, a questionnaire, followed by debrief in which the users will state their feedback on the SAFETY4RAILS Tool and the NGT, as laid out in Chapter 3.

4.2 Exercise 2 (Ankara, Turkey)

4.2.1 Objective of the exercise

The objective of the simulation exercise that will be held in Ankara is to evaluate both the first version of the S4RIS platform, in the context of a terrorist attack combining the explosion of a luggage abandoned and the intrusion of the terrorist inside an important room to perform a cyber-attack, and the individual tool capacities. The scenario is described in detail in D8.2.

The simulation exercises will involve several S4RIS capabilities to cover each resilience stage in the context of scenario: prevention, detection, response, recovery.

In this simulation exercise, 12 tools will provide some of their capabilities. Some of the tools will be integrated in the S4RIS platform whereas some others will be stand-alone. Full integration in the S4RIS platform is planned to be completed by the next simulation exercises.

4.2.2 What is going to be evaluated?

This section lays out the specific requirements that will be evaluated in this scenario. As laid out in Chapter 3, the organisation of the exercise will also be evaluated and all evaluation processes executed as specified.

4.2.2.1 Requirements for the Prevention phase

The functional requirements identified in SAFETY4RAILS Deliverable D1.4 and the corresponding objectives in the scenario that have been identified in SAFETY4RAILS Deliverable D8.2 are the following:

TABLE 7 EGO SIMULATION EXERCISE - REQUIREMENTS FOR THE PREVENTION PHASE

No	***Req.-ID - from D1.4***	**Short name**	**EGO Scenario objectives**	**Inte-grated in S4RIS (Y/N)**
EGO-PRE-1	CaESAR_02	CaESAR should identify weak points in the railway/metro system	Identification of the weakest/most critical components and associated cascading effects and overall resilience analysis of the infrastructure before the event.	Y
EGO-PRE-2	CAMS_02	Maintenance and repair budget calculation	The individual tool will be used to inform the metro operator on the budget to allocate to repair/maintain/rehabilitate the infrastructure after a set of possible events, therefore providing the necessary input to make a proactive plan and be ready in case of an attack. The metro operator will be also provided with information regarding the asset condition and degradation due to normal ageing, enabling timely response ahead of malfunctioning.	Y
EGO-PRE-3	DATA FAN-2	High prediction performance of results, e.g. anomaly detection	Provide information about the expected number of passengers early in the morning at the station where the attack takes place. The end-user will be able to run what-if scenarios to analyse how they will affect the number of passengers and delays in the infrastructure (e.g. the closure of a station).	Y
EGO-PRE-4	iCrowd_02	Simulate an evacuation because of terrorism (bomb, gas	Provide simulation capabilities to understand the probability of detecting a malicious actor attempting to break into the important room, therefore assessing the effectiveness of CCTV	*Y*

No	*Req.-ID - from D1.4*	Short name	EGO Scenario objectives	Inte-grated in S4RIS (Y/N)
		release) or natural disaster (fire/flood)	camera locations and guards, so that these could be eventually improved.	
EGO-PRE-6	PRIGM_01	Hardware encryption and random number generator modules	Cryptographic capabilities to protect critical data subject to vulnerabilities, truly random number generators for secret generation and secure One-time-password.	
EGO-PRE-5	PRIGM_06	Operations must be GDPR compliant	Apply data security and management policy and GDPR regulations (including both node and person authentication) so as to assure end-to-end security and improvement of the cyber resilience by preventing hardware-level attacks.	Y
EGO-PRE-7	RAM2_01	RAM2 should provide risk assessment and prioritisation	Provide vulnerability and security gaps assessment, along with risk assessment for each of the operational units in the metro system.	Y
EGO-PRE-8	SECURAIL_3	Computation of Risk	Enable off-line risk analysis of the metro infrastructure to understand the level of risk for each critical asset during a given hazardous event.	Y
EGO-PRE-9	SECURAIL_6	Cost-Benefit Analysis	Perform a cost benefit analysis of the infrastructure to understand which of the solutions analysed to reduce the risk level is the best, taking into account both costs and benefits.	
EGO-PRE-10	TISAIL_2	Detection of cyber-threats related to the railway sector: Internet-Exposed Assets and credential leaks	Provide situational awareness about vulnerabilities that could be exploited by attackers: e.g. CCTV, Power Grid, Windows 10.	Y

4.2.2.2 Requirements for the detection phase

The functional requirements identified in D1.4 and the corresponding objectives in the scenario that have been identified in D8.2 are the following:

TABLE 8 EGO SIMULATION EXERCISE - REQUIREMENTS FOR THE DETECTION PHASE

No	*Req.-ID - Need.-ID in D1.4*	Short name	Objective for the EGO exercise	Inte-grated (Y/N)
EGO-DET-1	CuriX_01	Anomaly detection (univariate and multivariate)	Evaluate how passenger flows correlate to each other, so to enhance/optimise the cascading effects analysis perform by the other S4RIS tools. Identify anomalies in passenger flows of other connected stations.	Y
EGO-DET-2	DATA FAN-7	Manner of the applied anomaly detection	Data gathered regarding the flow of passengers will be used to detect significantly high passenger volumes in stations and trains, particularly on days with really crowded events.	Y
EGO-DET-3	Ganimede_2	Enhanced abandoned baggage detection	Abandoned baggage detection based on video streaming from CCTV cameras.	Y
EGO-DET-4	PRIGM	*No correspondence found in D1.4*	Analysis of log data of main security operations (e.g. authentication, encryption, key exchange, etc) to determine anomalies, monitor the authentication flow for misuses/spoofing and help to discriminate between flooding data and normal flow. Furthermore, tracing and detection of cyber anomalies will be enabled, therefore assisting other countermeasure tools for enhanced resilience.	Y
EGO-DET-5	RAM2_02	RAM2 should generate correlated insights	Correlation of data gathered from multiple monitoring sources in order to detect potential threats. For example, it will be able to correlate the different attack vectors happening in the station.	Y
EGO-DET-6	SENSTATION_02	The resilience of the alternative secure data channel must be improved by end-to-end and hardware-based security.	Secure Gateway at edge nodes responsible from data protection where data is generated. It enables receive some instant information from the Electronic Equipment room to monitor unauthorised physical access, so that the operator can alert the security guard and the main Command and Control centre.	Y
EGO-DET-7	TISAIL_6	Integrate alerts related to cyber-threats in the railway sector with a MISP repository	Crisis Manager will be able to correlate the information (e.g., IoCs) provided by TISAIL for detecting threats in their networks using their security tools (e.g. IDS, SIEMs).	Y
EGO-DET-8	UNIMS_01	Unified management for networks, infrastructure and systems	Data monitoring (including network) offering capabilities for pre-configured critical conditions.	Y

4.2.2.3 Requirements for the response phase

The functional requirements identified in SAFETY4RAILS Deliverable D1.4 and the corresponding objectives in the scenario that have been identified in SAFETY4RAILS Deliverable D8.2 are the following:

TABLE 9 EGO SIMULATION EXERCISE - REQUIREMENTS FOR THE RESPONSE PHASE

No	*Req.-ID - Need.-ID*	Short name	Objective for the EGO exercise	Inte-grated (Y/N)
EGO - RES-1	CaESAR_05	Implementation and evaluation of mitigation measures	Support to end-user to select the appropriate mitigation measure to respond against different event in the Turkish infrastructure.	Y
EGO-RES-2	DATA FAN-2	High prediction performance of results, e.g. anomaly detection	Prediction of the number of passengers for a specific surrounding station in order to redistribute the passengers at the affected station.	Y
EGO-RES-3	iCrowd_05	Simulate access to a restricted area by cyber-attack (hackage of door) or physical attack (disabling a guard)	Know the time required to reach the important room in a case of emergency, which can be affected by the crowd congestion and the evacuation process, so to calculate the time for which the important room will be compromised and in an unknown state. Estimation of the total evacuation time and a distribution of evacuation times for passengers, so the performance of an evacuation plan can be assessed and improved.	Y
EGO-RES-4	RAM2_01	RAM2 should provide risk assessment and prioritisation	Risk-based prioritisation of issues, case management for tracking response actions. End user accesses the data through RAM2 Dashboard display. The user follows the prioritised alerts and mitigation steps for each of the alerts for risk reduction and response to detection of ongoing threats.	Y

4.2.2.4 Requirements for the recovery phase

The functional requirements identified in SAFETY4RAILS Deliverable D1.4 and the corresponding objectives in the scenario that have been identified in SAFETY4RAILS Deliverable D8.2 are the following:

TABLE 10 EGO SIMULATION EXERCISE - REQUIREMENTS FOR THE RECOVERY PHASE

No	*Req.-ID - Need.-ID*	Short name	Objective for the EGO exercise	Inte-grated (Y/N)
EGO-REC-1	CAMS_10	Assessment of recovery	Railway operator will be aware of vulnerability and fragility of the asset after the incident, so as to improve resource deployment and control financial loss in the future.	Y

4.2.3 Who is going to evaluate?

Overall, registration for the exercise is ongoing and the following is simply a preliminary list of evaluators' companies:

- All participants to the exercise
- End-user representatives organising the exercise:
  - EGO security team and other relevant departments.
  - TCDD security team and other relevant departments
- End-user representatives from the Consortium:
  - Rail partners: FGC, PRORAIL, RFI.
  - Metro partners: MDM.
  - Local authority: CDM.
- End-users' representatives outside the Consortium and mainly from the Advisory Board: a dedicated workshop will be organised for them.

4.2.4 How is it going to be evaluated?

The evaluation will be conducted through observations, a questionnaire, followed by debrief in which the users will state their feedback on the SAFETY4RAILS Tool and the NGT, as laid out in Chapter 3.

5. Conclusion

5.1 Main challenges and use of evaluation results

There are many challenges for the evaluation of the S4RIS platform by the end-users.

The first one is the scope of the S4RIS platform which is very broad: 18 tools addressing the 4 main stages of the resilience approach (prevention, detection, response, recovery). Therefore, the evaluation by the end-users can only be partial. A limited set of capabilities will be tested for each scenario given the timeframe for the simulation exercise which is quite short. The evaluation will highlight S4RIS novelties especially when dealing with combined cyber-physical threats.

The second one is related to the availability of data. Many tools are based on machine learning and need to be trained within a long time period to be even more efficient. Moreover, railways have not experienced many cyber-physical threats for now and therefore historical data is not available.

The evaluation of each exercise will serve to improve the tools and their integration as well as the outputs of the S4RIS platform for future exercises.

This evaluation will contribute to the assessment of the Technology Readiness level especially for TRL 6 (System/subsystem model or prototype demonstration in a relevant environment) and TRL 7 (System prototype demonstration in an operational environment). It will be used to assess whether the demonstrations/exercises have been performed successfully in a relevant environment. It will also be input into steps after the project to implement the results such as validation of products following also the UK FSR guidelines even more comprehensively.

5.2 Future work

The methodology described in this deliverable will served as a guide within task 8.3 on “Evaluation – End-user and developer feedback for improvement”. Within this task, some additional questions could be defined and included in the evaluation.

The methodology described in this deliverable is applied to the 2 first exercises (Madrid and Ankara). The scenario-based requirements evaluation for these 2 exercises are based on the information available in the draft deliverable D8.2 from 9 December 2021 and some adjustments are foreseen when the final version will be available. Therefore, the methodology will be further adapted progressively according to the progress of the next period when preparing the exercises.

Furthermore, it will be adapted for the two last exercises (Rome and Milano) to take into account both the results of the evaluation of the 2 first exercises and the specific requirements of the two last exercises that will be further described in a second stage (deliverable D8.3).

6. BIBLIOGRAPHY

Assila, A., Oliveira, K.M., & Ezzedine, H. (2016). Standardized Usability Questionnaires: Features and Quality Focus. Computer Science and Information Technology, 6.

Australian Disaster Resilience Handbook Collection, "Handbook 3 Managing Exercises", 2012.

Bevan N. International Standards for HCI. In Claude G , ed. Encyclopedia of Human Computer Interaction. Hershey, PA, USA: IGI Global ; 2006: 362 – 372.

Bevan N. International standards for usability should be more widely used. J Usability Stud. 2009;4(3):106–113.

Borland P (2000) Experimental components for the evaluation of interactive information retrieval systems. Journal of Documentation 56(1):71-90.

Brooke, J., 1996. SUS: A ‘‘quick and dirty” usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, A.L. (Eds.), Usability Evaluation in Industry. Taylor and Francis, London.

Chai, C. S., & Der-Thanq, V. C. (2004). A review on usability evaluation methods for instructional multimedia: an analytical framework. International Journal of Instructional Media, 31(3), 229.

Crabbe S., Lucas M., Ramm R., Sgrenzaroli M., Developmental Validation of the 3D-Forensics system, Version 1.1, September 2020.

Craig, R. Jaskiel, S. « Systematic Software Testing. » Artech House, 2002.

Department of Housing, Local Government and Heritage, "A framework for major emergency management", September 2021.

Emergency telecommunications table-top simulation guide, 2020.

European Commission, HORIZON 2020 – WORK PROGRAMME 2014-2015 General Annexes, G. Technology readiness levels (TRL).

Hsin-Ke, L., Sung-Chun, T., Peng-Chun, L., Kuo-Chung, C., & Chen, A. N. (2020). Toward a new real-time approach for group consensus: A usability analysis of synchronous Delphi system. Group Decision and Negotiation, 29(2), 345-370; CAN, Gülin Feryal; DEMIROK, Seda. Universal usability evaluation by using an integrated fuzzy multi criteria decision making approach. International Journal of Intelligent Computing and Cybernetics, 2019. Dawood, K. A., Sharif, K. Y., Ghani, A. A., Zulzalil, H., Zaidan, A. A., & Zaidan, B. B. (2021). Towards a unified criteria model for usability evaluation in the context of open-source software based on a fuzzy Delphi method. Information and Software Technology, 130, 106453.

International Laboratory Accreditation Cooperation, Modules in a Forensic Science Process.

International Organization for Standardization. Ergonomic requirements for office work with visual display terminals (VDTs) Part 11 Guidance on usability. International standard, 9241-11. Geneve: ISO; 1998.

International Organization for Standardization. Ergonomics of human-system interaction - Part 210: human-centred design for interactive systems. International standard, 9241-210. Geneve: ISO; 2010.

International Organization for Standardization Ergonomics of human-system interaction: part 11: usability: definitions and concepts (ISO/DIS 9241-11.2:2016).German and English version prEN ISO 9241-11:2016.

ITU (International Telecommunication Union) Publications, "Handbook Evaluation of Exercises", Swedish Civil Contingencies Agency, March 2011.

J. Pries-Heje, R. Baskerville, and J.R. Venable, Strategies for Design Science Research Evaluation. In: Proceedings of the 16th European Conference on Information Systems (ECIS 2008), Galway, Ireland, 2008.

James R. Lewis (2018): Measuring Perceived Usability: The CSUQ, SUS, and UMUX, International Journal of Human–Computer Interaction, DOI:10.1080/10447318.2017.1418805

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction, 7, 57–78.

Lewis, J.R. (2018). Measuring Perceived Usability: The CSUQ, SUS, and UMUX. International Journal of Human–Computer Interaction, 34, 1148 - 1156.

SAFETY4RAILS Deliverable D1.1 (October 2020).

SAFETY4RAILS Deliverable D1.4 (October 2020).

SAFETY4RAILS Draft Deliverable D8.2 from 09/12/2021.

SAFETY4RAILS Grant Agreement, version 2.0 with amendment (June 2020).

Sarah Brown (2007), Standardized Technology Evaluation Process (STEP) User's Guide and Methodology for Evaluation Teams.

Sauro, J., & Lewis, J. R. (2016). Quantifying the user experience: Practical statistics for user research (2nd ed.). Cambridge, MA: Morgan Kaufmann.

Tasmania Fore Service, "Tasmanian Exercise Framework Evaluation Plan Exercise Transfundo", May 2013.

UK Forensic Science Regulator, Codes of Practice and Conduct For Forensic Science Providers and Practitioners in the Criminal Justice System (FSR, Issue 7, 2021).

UK Forensic Science Regulator, Guidance: Validation (FSR, Issue 1, 2014).

UK Forensic Science Regulator, Guidance: Validation (FSR, Issue 2, 2020).

UK Forensic Science Regulator, Method Validation in Digital Forensics, (FSR, Issue 2, 2020).

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: a Framework for Evaluation in Design Science Research. European Journal of Information Systems, 25, 77-89.

Vertesi, A., Dogan, H., & Stefanidis, A. (2020). Usability Evaluation of Virtual Learning Environments: A University Case Study. In Online Teaching and Learning in Higher Education (pp. 161-183). Springer, Cham.

Vom Brocke J., Hevner A., Maedche A. (2020) Introduction to Design Science Research. In: vom Brocke J., Hevner A., Maedche A. (eds) Design Science Research. Cases, Cham. https://doi.org/10.1007/978-3-030-46781-4_1.

Zafiropoulos K. (2005): How to write down a scientific paper. Athens. Kritiki.

7. ANNEXES

7.1 ANNEX I. Glossary and Acronyms

TABLE 11 GLOSSARY AND ACRONYMS

Term	Definition/description
AB	Advisory Board
CCTV	Closed-circuit television
CDM	Comune di Milano
CO	Confidential
CSUQ	Computer System Usability Questionnaire
D	Deliverable
DC	Data controller
DoA	Description of the Action (Annex 1 to the Grant Agreement)
DSRM	Design Science Research Methodology
EC	European Commission
EER	Electronics Equipment
ENISA	European Union Agency for Cybersecurity
FEDS	Framework for Evaluation in Design Science
FGC	Ferrocarrils de la Generalitat de Catalunya
GUI	Graphical User Interface
IDS	Intrusion Detection System
InfoQual	Information Quality
IoCs	Input Output Control Systems
IPR	Intellectual Property Rights
ISO	International Standardisation Organisation
ITU	International Telecommunications Union
KPIs	Key Performance Indicators
LEA	Law Enforcement Agency
MDM	Metro de Madrid
MISP	Malware Information Sharing Platform
NGT	Nominal Group Technique
OAJR	Observation, Analysis, Judgement, Recommendations (assessment criteria)
PSIM	Physical security information management
PSSUQ	Post-Study System Usability Questionnaire
PTZ cameras	Pan, Tilt and Zoom cameras
Req	Requirement

S4RIS	SAFETY4RAILS Information System
SIEMs	Security information and event management
SysUse	System Usefulness
TC	Technical Committee
TOC	Train Operating Company
TRL	Technology Readiness level
TSI	Technical Specification for Interoperability
UC	Use-Case
UR	User Requirement
WG	Working Group
WP	Work-Package
WS	Workshop

7.2 ANNEX II. End-users' evaluation methodologies

7.2.1 CDM methodology

The Milan Municipality (Comune di Milano) needs to consider a complex network, with many different actors each one with different tasks and responsibilities, while an emergency plan is built on different stakeholders and on communication among many different nodes and touchpoints.

That is why it is not easy to identify a unique methodology, but it is necessary to refer to the different methodologies used by each actor, without losing however the complexity of the scenario.

This said about methodology, the definition of some KPIs outlined indicators that could fit a complex interaction as "integrated" indicators of a process that will deploy different tools and involve different actors: thinking about the main features of the pilot exercise: service security, big amounts of people moving at the same time, cascade effects, socio-economic impacts.

The KPI list should be discussed and validated with the tool providers and the project partnership, in order to fit with an expert-based evaluation.

Some proposed KPIs

- time of service recovery after an interruption/stop/accident
- n. of service interruptions due to extreme climate or meteorological events (severe interruptions)
- n. of service interruptions/delays due to extreme climate or meteorological events (light interruptions)
- n. of crowd gatherings due to service problems
- n. of official complaints (and type of complaints) due to service interruptions/delays
- amount of money estimated for service recovery/assurance/refunds/damages

7.2.2 FGC Methodology

Serving more than 90 million passengers a year, FGC has consolidated its position during its 40 years of activity around values such as security, trust and innovation. Consequently, the company aims to innovate continuously to ensure the security of its passengers, as well as to maintain a high level of trust from them.

Therefore, to ensure that the changes proposed in the company improve passenger security while maintaining a high level of passenger confidence in FGC services, the following procedures have been established:

- Set up of KPIs reflecting the most important aspects of FGC passenger security.
- A standardised change management procedure methodology to analyse the possible consequences of a change introduced in the operation of the company and for having a historical track of changes to follow and assess them.

Security KPIs
The following Table encompasses the different security KPIs considered by FGC when evaluating security solutions:

Category	KPI
Conduct linked to persons	Insults and threats: to agents and passengers
	Aggressions: to agents and passengers
	Gender violence
Robberies and thefts	Robberies
	Thefts
Damage	Stones thrown at trains
	Graffiti at stations

Category	KPI
	Area of graffiti at stations (m^2)
	Damage to installations
Attacks by gangs of graffiti vandals	Attacks: avoided, completed, total
	Number of graffities in protected and unprotected zones
	Area of graffiti at trains (m^2)
	Events with trapped graffiti gangs
Antisocial conducts	Altercations and inconvenience to customers
	Jumping/crossing of the tracks
	Misuse of alarms / unlocking of doors on trains
	Other antisocial conducts
	Total antisocial conducts
	Complaints lodged
Incivism alerts (App)	Number of alerts
	Incivism alerts managed (%)

Change Management Procedure Methodology

The FGC Change Management Procedure is based on the requirements of European Regulation 402/2013 on the adoption of a common safety method for risk evaluation and assessment, as amended by Regulation (EC) No 2015/1136 of the European Parliament and of the Council amending Implementing Regulation 402/2013.

Through the aforementioned regulation, the European Commission has introduced the rule that in an existing railway system, such as that of FGC, any change that may have a significant impact on railway safety and security must be accompanied by a risk management procedure. Changes include both technical changes (e.g., the creation or extension of a line, the introduction of a new signalling system, etc.) as well as changes in operating or maintenance processes (e.g. the modification of a maintenance plan), organisational changes, etc.

The regulation sets out the following key concepts:

- A list of criteria to be considered when deciding whether a change is significant from a safety and security point of view.
- A process to be followed for managing the risks that the change could introduce, based on a risk analysis and the creation of a register documenting the risks identified and how they are managed,
- The scope of the mission of an Assessment Body which has to give an opinion on the adequacy of the risk management carried out in accordance with the requirements of the Regulation. This body must be independent and accredited.

Each FGC department is responsible for managing the risks involved in the changes it promotes in the FGC railway system, analysing whether the changes are significant and documenting them and whether they are carrying out risk analysis and management procedures, as well as their evaluation from an external body.

Thus, the different departments of FGC establish the means and responsibilities within their area for the implementation of this procedure.

The Methodology (Figure 4) is divided into three subprocesses that are described below. It should be noted that work related to Risk Analysis and Risk Management may be carried out internally to FGC or may be subcontracted, in part or in full.

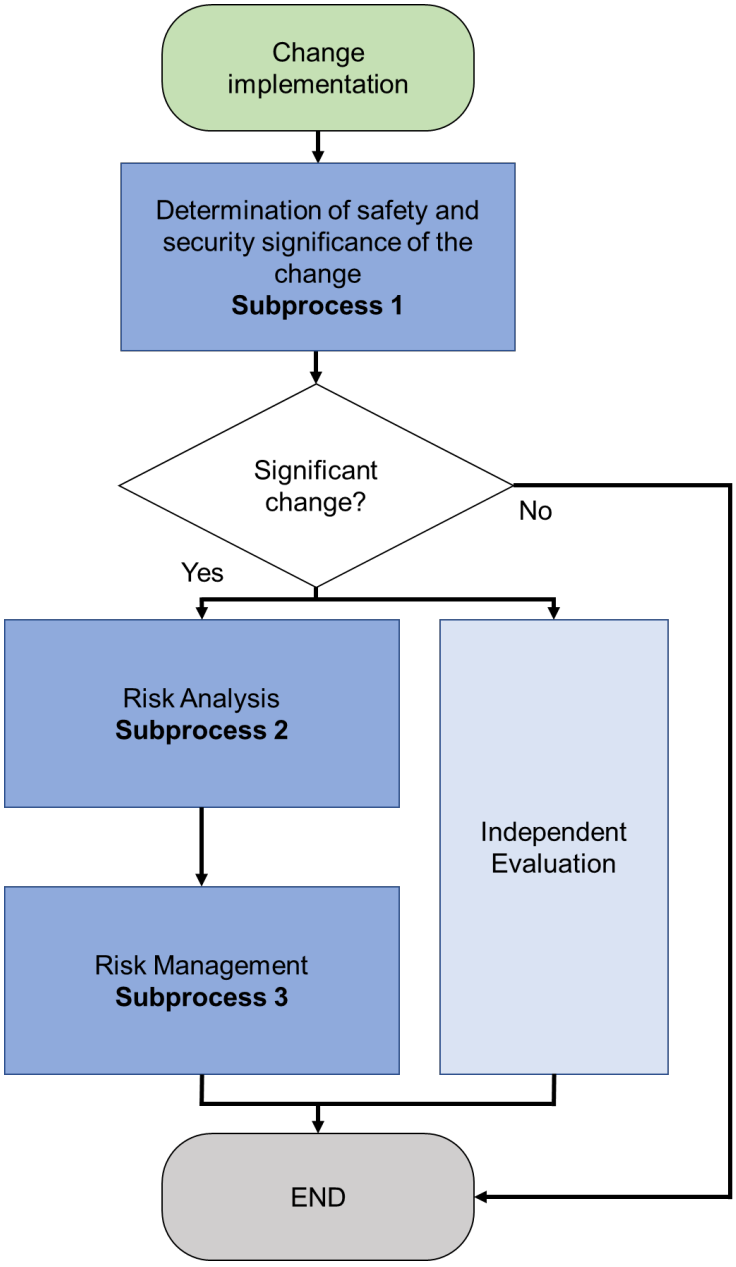


FIGURE 4 FGC CHANGE MANAGEMENT METHODOLOGY

- **Subprocess 1: Determination of safety and security significance of the change**

The conduct of risk assessment and risk management starts with the determination of the significance or non-significance of a change (in terms of safety).

The significance of the change is assessed by FGC using the process described in another flowchart that contains several forms to be completed. These include:

  - Form with the preliminary definition of the system that includes description of the change, description of the actual system, subsystems/components affected, FGC departments affected, indication of whether the change could have an impact on safety and security.
  - In the case that the change has an impact on safety and security, an evaluation of the significance is carried out by completing another form that takes the following aspects into account: consequences in case of failure, innovation used in the change realisation, change complexity, supervision ability, reversibility, additional measures etc.

From these forms, an Index of Significance is calculated. If this Index exceeds a predefined threshold, the change will be considered significant. If not, the change management procedure is ended. If the change is

considered significant, it will be necessary to proceed with Subprocesses 2 and 3 of the Change Management Procedure.

- **Subprocess 2: Risk Analysis**

The risk analysis is assessed by FGC using the process described in another flowchart that contains several forms to be completed. These forms include:

  - Potential accident, risk, possible causes, subsystem of the origin of the cause, importance, Risk Acceptance Principle, safety and security requirements etc.
  - Depending on the Risk Acceptance Principle, different procedures are required.

- **Subprocess 3: Risk Management**

Once the risks are identified and the safety and security requirements are defined by means of the Risk analysis, the Risk Management is assessed by FGC using the process described in another flowchart that contains a Risk Management form, which includes some parts of the Risk Analysis subprocess, as well as the following:

  - Security and Safety Requirement ID.
  - Security and Safety Requirements.
  - Security and Safety Measures.
  - Department responsible of the Risk Management.
  - References.
  - Status of the Requirement: it can be open, controlled, cancelled, transferred, closed.

Each operational area will describe the safety and security measures that enable compliance with the safety and security requirements and will transfer this information to the responsible department. The responsible department will update the status of each requirement accordingly. Likewise, the management of the registration of risks is carried out by the responsible department.

- **Independent Evaluation**

The Common Safety Methods Regulation provides that for each significant change, an independent evaluation by an Assessment Body shall be carried out for the correct application of the risk management process referred to the annexes and forms of the document, as well as for the results obtained. Hence, for each significant change, FGC will hire an independent Assessment Body. The Regulation requires the Assessment Body to be accredited.

The Assessment Body will:

  - Ensure that FGC fully understands the significant change from the documentation provided by the proposer.
  - Carry out an assessment of the safety, security and quality management processes followed during the design and implementation of the significant change.
  - Carry out an assessment of the implementation of the safety, security and quality processes during the design and implementation of the significant change.

Once the evaluation has been completed, the Assessment Body will deliver the Independent Evaluation report, which includes the following content:

  - Identification of the Assessment Body.
  - Independent Assessment plan.
  - The definition of the scope covered by the independent assessment and its limitations.
  - The results of the independent evaluation activities carried out to verify the compliance with the provisions of Regulation 402/2013.
  - Any cases of non-compliance, as could be detected, with the provisions of the Regulation 402/2013 or with the recommendations of the assessment body.

- Conclusions of the independent assessment.

Once the Independent Evaluation and the Risk Management subprocess have been completed, each operational area is responsible to apply the safety and security measures defined, or, if not responsible, transfers the measures to the responsible area. Furthermore, the responsible area will update the status of each requirement accordingly.

7.2.3 ProRail Methodology

Within ProRail, evaluation methodology is based on statistics. A safety dashboard with security KPIs for each kind of threat is available.

The type of threats are the following:

- Throwing objects at a train.
- Bomb/suspicious object alert.
- Arson.
- Graffiti.
- Copper theft (possible/attempt).
- Track circuit interrupted (for instance with a coin).
- Vandalism/destruction.

For each type of threat, the numbers below are reported:

- Total number in previous year.
- Norm for previous year (total).
- Total number in current year until the specific month (cumulative).
- Norm in the current year until the specific month (cumulative).
- Norm for current year (total).

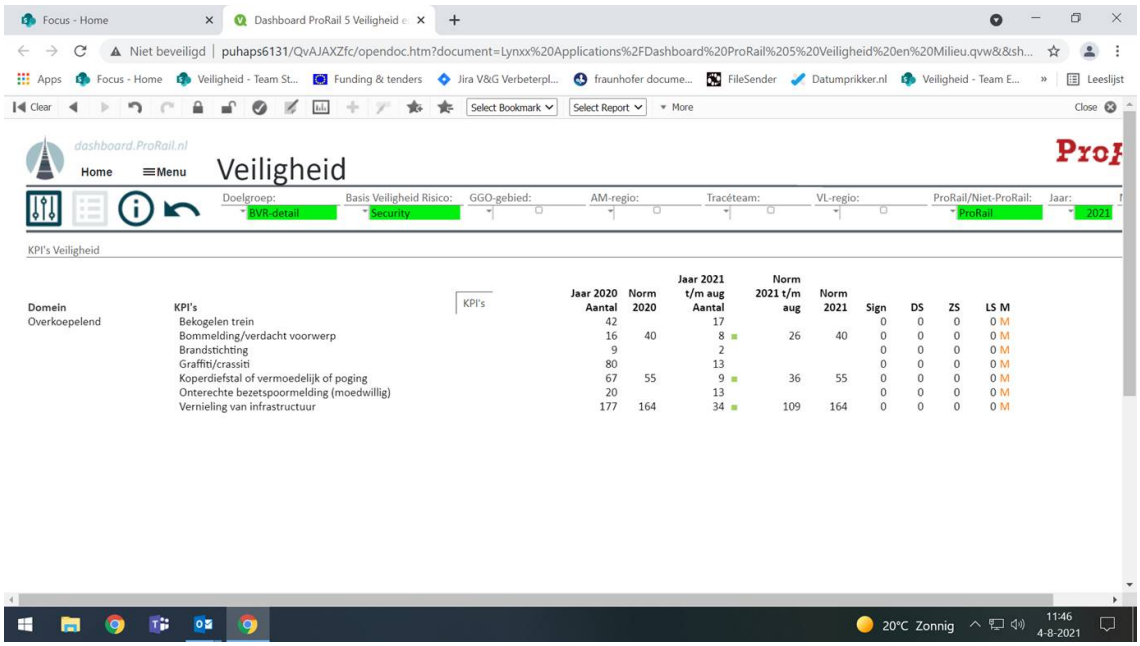


FIGURE 5 PRORAIL - SECURITY INCIDENT STATISTICS

7.2.4 Polish State Railways (PKP S.A.) methodology

There is no existing paper evaluation methodology implemented in the company. The evaluation of each solution is based on employees' knowledge and experience and conducted accordingly. However, new technical security systems installed at train stations have to meet the following criteria:

a) CCTV:
   - IP based.
   - Client-server architecture.
   - Should integrate tools from different providers into one system.
   - Quality requirements (number of pixels per meter).
   - Minimum resolutions of cameras.
   - Depending on the station (if justified) it is also allowed to use analytics in order to
     - Count people along with the direction of their movement and defining routes of people's movement.
     - Create heat maps.
     - Integrate with the lighting control system (possibility of making data available for lighting control).
     - Plate recognition.
     - Other types of image analysis.
   - Storing period.
   - Minimum parameters for:
     - Indoor/outdoor PTZ cameras.
     - Indoor/outdoor fixed cameras.
     - Micro-bullet cameras.
   - CCTV recorders.
   - Video management system.
   - Video anonymization software.

b) Access control system:
   - IP based.
   - Client-server architecture.
   - Database should be based on SQL.
   - Should be able to be integrated with PSIM.
   - Detailed log of all events and alarms is required.
   - Automatic database backups and copies of the full system configuration should be provided.
   - Access control readers.
   - Cards should be equipped with standard communication protocols.
   - Access control system card programmer.
   - Standard interfaces of control panel and controllers.

c) Burglar alarm system:
   - In accordance with PN-EN50131 norm.
   - To able to be integrated with PSIM.
   - To define privilege profiles.
   - Search tool allowing to check all events.

SAFETY4RAILS

Partners:

Fraunhofer EMI
UIC
Metro de Madrid
EGO Genel Müdürlüğü
RFI Rete Ferroviaria Italiana – Gruppo Ferrovie dello Stato Italiane
TCDD
Leonardo
ceis avisa partners
STAM Mastering Excellence
information company
RMIT University
University of Reading
etra I+D
Demokritos National Centre for Scientific Research
RINA
WINGS ICT Solutions
Newcastle University
EOS European Organisation for Security
Innova Integra
LAUREA University of Applied Sciences
CybeServices Made in Europe
erGE
FGC Ferrocarrils de la Generalitat de Catalunya
MTRS
INTRACOM TELECOM
Universitas Miguel Hernández
TREE Technology
Elbit Systems C4I and Cyber
ProRail
Comune di Milano

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 883532.