Horizon 2020
Communications Networks, Content & Technology
Cloud & Software
Research & Innovation Action
Grant agreement n° 731529

Title:	WP5 – D5.2 – Validation Roadmap and framework V1
Date:	August 31, 2017
Writer:	Jesús Gorroñogoitia Cruz - Atos Caleb James DeLisle - XWiki Lars Thomas Boye – TellU Stéphane Laurière – OW2 Mael Audren - ActiveEon
Reviewers:	Lars Thomas Boye, Caroline Landry, Benoit Baudry

Table Of Content

1. Executive Summary

This document describes the objectives of the validation activities, the proposed evaluation method, the supporting conceptual and material evaluation framework and the adopted evaluation roadmap, which altogether drive and support the industrial validation of the STAMP toolset and services. The method collects and defines a set of configurable process fragments or tasks that, once tiled together, constitute a customize methodology for conducting the different validation activities. It is influenced by industrial standards and scientific guidelines to conduct and report assessment experiments in real industrial scenarios. The framework introduces the key evaluation concepts and elements, but also the supporting materials and tools required to conduct the validation activities following the proposed method. The roadmap schedules a set of proposed validation activities in a timeline within the project lifetime, which is aligned to the STAMP development roadmap, milestones and release plan, aiming at maximizing the impact of the evaluation feedback (e.g. findings, recommendations, reported issues, etc.), reported by stakeholders, into the quality of the STAMP toolset and services.

This document provides the first conceptualization for the evaluation method, framework and roadmap. It will be adopted during the following M9-M20 evaluation period. During this period, this conceptualization will be further refined and instantiated. The final agreed instantiation will be formalized in the following release of this document, D5.4, due by M20.

2. Revision History

Date	Version	Author	Comments
30-jun-2017	0.1	Jesús Gorroñogoitia (Atos)	ToC
05-jul-2017	0.2	Jesús Gorroñogoitia (Atos)	Initial contributions (sections 7, 8, 10)
10-jul-2017	0.3	Caleb James DeLisle (XWiki) Lars Thomas Boye (TellU)	Revision and contributions to section 7.1 Revision and contributions to section 8 Section 3 and 4
17-jul-2017	0.4	Jesús Gorroñogoitia (Atos)	Section 11 Evaluation Method
17-jul-2017	0.5	Jesús Gorroñogoitia (Atos)	Template for issue reporting
20-jul-2017	0.6	Jesús Gorroñogoitia (Atos)	Peer-review ready version
08-Aug-17	0.7	Lars Thomas Boye (TellU)	Peer-review
22-Aug-17	0.8	Jesús Gorroñogoitia (Atos) Stéphane Laurière (OW2)	Addressing peer-review recommendation Contributions to section 10
30-Aug-17	0.81	Caroline Landry, Benoit Baudry (INRIA)	Peer-review
04-Sep-17	0.9	Jesús Gorroñogoitia (Atos)	Addressing peer-review recommendations
05-Sep-17	0.91	Caroline Landry (INRIA)	Peer-review
06-Sep-17	1.0	Jesús Gorroñogoitia (Atos)	Addressing peer-review recommendations Re-structure of document

3. Objectives

The objectives of this deliverable are two-fold. On the one hand, this deliverable defines a conceptual and practical methodology and its supporting framework that will be adopted to conduct the validation of the STAMP toolset and services in the context of the different industrial use cases (T5.3-T5.7) that participate in the project. This validation methodology will also define specific support for external validation by the potential adopters of the STAMP technology, among those developers of open communities who are interested on test amplification techniques and their potential benefits in software development and QA. This document elaborates on important concepts of the validation such as its objectives, participants, validation environments, validation activities, supporting materials, metrics and measurements or reporting.

On the other hand, this document defines a validation roadmap that schedules the validation activities in a timeline, in a suitable way for their alignment to the STAMP development activities and its plan for software releases. This roadmap tries to maximize the impact that the validation findings may have on the improvement of the quality and usability of the STAMP tools, particularly assisting the team during their development lifecycle, as well as facilitating their adoption by practitioners external to the project consortium.

4. Introduction

This document defines the methodology and supporting framework adopted to conduct the validation of the STAMP techniques and its toolset in real-life industrial scenarios. Software validation is an important part of the software development life-cycle, sometimes mistaken with a related concept, software verification, which is not targeted in this document:

- *Software validation* assesses whether or not a software product satisfies or fits the intended use, that is, the software meets the stakeholders' requirements. In other words, "developers have built the right system".
- *Software verification* assesses whether or not the designed and developed software is compliant to its specification, that is, "developers have built the system right". The verification of the STAMP toolset will be assessed by the development teams themselves and reported in the deliverables associated to the STAMP individual tools.

In summary, the **validation** of the STAMP toolset and services is the main scope of the methodology and framework described in this deliverable.

The validation activities adopted will combine two different methodological approaches:

- A validation conducted on continuous basis, adopting the STAMP toolset and services on the daily testing activities of the development life-cycle of the industrial software development projects included in the use cases. These validation activities will report feedback to STAMP development teams in continuous basis as well.
- A validation conducted on discrete basis, assessing the STAMP toolset and services in scheduled workshops conducted on laboratory controlled experimentation.

Besides the continuous validation feedback reported on continuous basis to the STAMP development teams, additional validation findings will be formally reported in official deliverables D5.5, D5.6 and D5.7

The decision of combining both validation approaches is justified to take advantage of the benefits of both approaches: a) influencing the STAMP toolset functional and usability specification and technical development with its duly validation on real testing phases of the industrial software development life-cycle and b) assessing the hypotheses associated to the proposed KPIs in controlled laboratory experimentation.

The remaining of this document is structured as follows. Section 5 collects the references cited inline the document. Section 6 collects the abbreviations used within the documents. Section 7 explains the objectives of the STAMP validation process. Section 8 describes the validation roadmap and its alignment to the STAMP software release plan. Section 9 concludes the document. Section 0 introduces all the methodological concepts and constituents of the validation framework. Section 11 introduces the elements of the adopted evaluation method.

5. References

[0] J. Gorroñogoitia et al. D5.3 Validation Roadmap and framework. V1, STAMP report. 2017. Latest version available at:

https://gitlab.ow2.org/stamp/h2020/blob/master/deliverables/wp5/d52_validation_roadmap_and_framework.pdf

[1] M. R. Fine, Beta Testing for Better Software, ISBN 0-471 -25037-6, Wiley Computer Publishing, 2002

[2] Z. Yanga, S. Cai, Z. Zhou, N. Zhou, Development and validation of an instrument to measure user perceived service quality of information presenting Web portals, Information and Management, Elsevier, 2004

[3] J. Rubin, D. Chisnell, Handbook of Usability Testing - How to Plan, Design, and Conduct Effective Tests, ISBN: 978-0-470-18548-3, Wiley Publishing, 2008

[4] E. Almirall, M. Lee, J. Wareham, “Mapping Living Labs in the Landscape of Innovation Methodologies”, Technology Innovation Management Review, 2012

[5] Barry W. Boehm, “Characteristics of Software Quality”, North-Holland Pub. Co., 1978

[6] Jedlitschka, A., & Pfahl, D. (2005, November). Reporting guidelines for controlled experiments in software engineering. In Empirical Software Engineering, 2005. 2005 International Symposium on (pp. 10-pp). IEEE.

[7] C.J. DeLisle et al. D5.1 Industrial requirements and metrics for validation. V1, STAMP report. 2017

[8] Cohn, Mike. User stories applied: For agile software development. Addison-Wesley Professional, 2004.

[9] Rougemaille, Sylvain, et al. "Methodology fragments definition in SPEM for designing adaptive methodology: A first step." International Workshop on Agent-Oriented Software Engineering. Springer, Berlin, Heidelberg, 2008

[10] Björn Regnell, Richard Berntsson Svensson, and Thomas Olsson, “Supporting roadmapping of quality requirements”, Software, IEEE 25.2(2008): 42-47, 2008;

6. Acronyms

D<X.X>	Deliverable <X.X>
GA	Grant Agreement
EC	European Commission
KPI	Key Performance Indicators
M<x>	Month <X>
QA	Quality assurance
UC	Use Case

7. Validation Objectives

The overall goal of the validation task is to conduct a fit-for-purpose evaluation that assesses the STAMP toolset and services utility in real life scenarios that requires test amplification. The aim is to assess whether the STAMP scientific and technical achievements satisfy the needs and expectations of users.

The results of the project will be evaluated in the context of real life situations, starting with selected industrial use cases (UCs) owned by partners of the consortium, but also concluding with the involvement of communities of open-source software developers that will be invited to use and evaluate the STAMP toolset and tools in their own development activities. For this purpose STAMP will provide public downloadable releases of the standalone toolset and eventually links to publicly accessible instances of some services[1] – in the last phase of the project –, so potential adopters within those communities can download the tools (or eventually access the services) and evaluate them.

The approach used in the STAMP project is not defined entirely from scratch. Instead, partners involved in the evaluation consider and reuse parts of existing research and industry methodologies, methods, practices and standards related to requirements validation, requirements reviews, acceptance testing, beta testing [1], user-perceived service quality [2] and usability [3], considering their applicability for STAMP validation purposes. Some shared characteristics of participatory validation observed in the living lab methods [4] are derived, in particular the shared characteristic to involve users early in the process of validation in real-life environments and to incorporate the validation results into the toolset and service development.

The user needs are formalized as a set of requirements (and associated KPIs designed to assess their fulfillment). T5.1 “coordinates the elicitation of these industrial requirements for test amplification from the use case stakeholders, aiming at influencing the scientific and technical development of WP1-WP4 and the selection of results for industrial exploitation in WP6.”

As explained in D5.1 [7] and the GA, the final validation of the project will come through the validation of the objectives and KPIs. Evolution of the project as a whole will be validated through repeated collection of the KPI metrics by the use cases as they integrate the STAMP software and methodologies. In order to target the specific issues that are needed to improve the use case KPIs, we have developed a requirements and validation framework to allow use cases to express their needs clearly to the scientific and development work-packages.

Aiming at making this document focused the detailed description of the adopted evaluation framework and method has been placed in the appendices. Elements of the evaluation framework are introduced in Appendix A, section 0. Elements of the adopted evaluation method are introduced in Appendix B, section 11.

[1] STAMP Evaluation will focus on the evaluation of the STAMP standalone tools, setting aside the evaluation of the services to the moment they are available.

8. Validation Roadmap

8.1. Alignment to the STAMP Toolset Release plan

This section introduces the timeline of validation activities that will be conducted along the project lifetime, structured in a number of phases, which are aligned to the STAMP tool and service release plan (see Table 1). According to the STAMP GA, a number of tool and service releases will made available according to the following schedule:

- STAMP Initial prototype on M12-M14 (November 17-January 18);
- STAMP Enhanced prototype on M20-M24 (July 18 - November 18);
- STAMP Consolidated tool on M34 (Sep 19);
- STAMP Final services on M36 (Nov 19).

Table 1 STAMP Tool and Service releases

Release	Date/Milestone	Deliverables (Responsible)
Initial prototype	M12/MS8 M14	D12 Initial prototype of the unit test amplification tool (INRIA) D22 Initial prototype on configuration test amplification (SINTEF) D32 Initial prototype of log optimization tool (TUD) D42 First public version of the API and initial implementation of services (ENG)
Enhanced prototype	M20/MS12 M24	D13 Enhanced prototype of the unit test amplification tool (INRIA) D23 Enhanced prototype of the configuration amplification (SINTEF) D33 Prototype of amplification tool for common and anomaly behaviors (TUD) D43 Second public version of the API and consolidated implementation of services (ENG)
Consolidated tool	M34/MS15	D14 Consolidated tool for the unit test amplification, selection and execution (INRIA) D24 Consolidated tool for the configuration amplification, selection and execution (SINTEF) D34 Consolidated services for online-test amplification (TUD)
Final services	M36	D44 Final public version of the API and consolidated implementation of services (ENG)

Above Table 1 further identifies the dates (and milestone) of each release, associated STAMP deliverables and responsible partners. Note that first and second releases in the table aggregate the releases of the tools and services, which separately take place with 2 month span. This is done so, because the evaluation activities target the assessment of the complete toolset, not only core tools, but their industrialization with tool facades (e.g. Maven/Gradle, CMI, IDE clients, services) as well. The exception to this approach is the explicit split of releases: *Consolidate Tool* and *Final Services* at M34 (MS15) and M36. In this case, final evaluation activities target the *Consolidate Tool* release, and not the *Final Services,* since they are released at the end of the project. Nonetheless, intermediate releases of the *Final Services* will be evaluated during the final stages of the project.

Additionally, a number of early prototypes for some tools (e.g. Descartes PITest and DSpot, both in WP1) are available from the starting of the project, as they are provided as initial baseline.

8.2. Validation phases and timeline

The validation timeline is organized in such a way that we address the following validation objectives:

- The validation activities provide early and continuous feedback to the development teams in technical WP1-WP4, during the entire development lifecycle, starting as early as initial prototypes are available for assessment.
- The validation activities are organized in such a way that they provide iterative, increasing and agile feedback:
  - Iterative:
    - assessing new toolset releases (internal for the consortium or external for the public) as soon as they are published;
    - assessing new toolset minor versions, as soon as they are patched in reaction to communicated bugs, defects or requested features;
  - Incremental:
    - incrementing the assessment precision and/or complexity by providing deeper assessment in increasingly complex usage scenarios;
  - agile:
    - adapting the assessment process to the toolset maturity and the validation objectives agreed between tool developers and evaluators.

Observing these considerations, the validation timeline has been organized in 3 phases that are aligned to the STAMP tools and services release plan, in a way that it optimizes the feedback provided to the development team at the suitable time that it can usefully support the development of the tools.

- **Phase 1 (M6-M18):** focuses on validating the main concept and the functionality of the STAMP toolset, from the industrial point of view, engaging other potential industrial adopters, as well as on providing early feedback to support the development of a usable and stable toolset;
- **Phase 2 (M19-M30)**: focuses on validating the first stable STAMP toolset, focusing of their usability, effectiveness, efficiency and the perceived fit-for-a-purpose benefit in a variety of industrial real scenarios. It also starts conducting initial in-lab controlled experiments to verify the fulfillment of the hypothesis claimed in KPIs;
- **Phase 3 (M31-M36)**: focuses on engaging open communities of developers for reporting feedback on the usage of the STAMP toolset on common daily development and testing activities. It also conducts extensive In-Lab controlled studies aiming at verifying the fulfillment of the hypothesis and claims associated to the baseline of KPI described in the

GA. Eventually, STAMP KPIs can be compared to those KPIs measured when an industrial case adopts any previous state of practice technique that targets functional challenges similar to STAMP ones.

Next Table 2 lists the breakdown of evaluation activities included in the roadmap, structured according to the different phases. Each activity is positioned into the STAMP project calendar. Activity coordinator and contributors are also specified. The associated official deliverable for reporting the evaluation activities and findings is identified. See section 10.7 for a description of the validation activity types mentioned in the *Evaluation Activity* column.

Table 2 STAMP evaluation roadmap: breakdown of activities

Evaluation Activity	**Objectives, Tool Target**	**Calendar**	**Coordinator**	**Contributors**	**Associated Reporting Deliverables**
Phase 1	Main STAMP concept and functionality validation. Early feedback on tool prototypes **Tool Target**: Initial Prototype (D12, D22, D32, D42)	M6-M18	Aeon	Atos, TellU, XWiki, OW2	D5.5 UC Validation Report V1
Pilot Trial	Reporting usage experiences with initial prototypes: issues and potential bugs. Assessment of concrete requirements and KPIs	M6-M18	Aeon	Atos, TellU, XWiki, OW2	
User Survey	Collect feedback about STAMP approach, concept, toolset functionality	M10-M13	Atos	AEon, TellU, XWiki, OW2	
Focus Group	Refinement, formalization and prioritization of STAMP requirements and KPIs	M14-M18	XWiki	AEon, Atos, TellU, OW2	
Phase 2	Usability, effectiveness, efficiency and fit-for-purpose validation. Initial verification of KPIs' hypothesis **Tool Target**: Enhanced Prototype (D13, D23, D33, D43)	M19-M30	OW2	Aeon, Atos, TellU, XWiki	D5.6 UC Validation Report V2
Pilot Trial	Reporting usage experiences with enhanced tools: issues and potential bugs. Early assessment of concrete requirements and KPIs	M19-M30	OW2	Aeon, Atos, TellU, XWiki	
In-lab controlled experiments	Assess the efficacy and efficiency of the STAMP enhanced tools in tasks conducted in common industrial scenarios. Initial assessment of the hypothesis	M25-M30	XWIKI	Aeon, Atos, TellU, OW2	

	associated to the baseline of KPIs				
Phase 3	External validation from open communities of developers. Final verification of KPIs' hypothesis. **Tool Target**: Consolidated Tool (D14, D24, D34), Final Services (D44)	M31-M36	Atos	AEon, TellU, XWiki, OW2	D5.7 UC Validation Report V2
Open Field Trial	Evaluate the STAMP concepts, methods and tools in daily practical testing situations dealt with by developers engaged in open-source development communities	M31-M36	OW2	Aeon, Atos, TellU, XWiki	
In-lab controlled experiments	Assess the efficacy and efficiency of the STAMP final tools in tasks conducted in common industrial scenarios. Final assessment of the hypothesis associated to the baseline of KPIs	M35-M36	Atos	AEon, TellU, XWiki, OW2	

Above evaluation activities will be conducted in the context of the following GA WP5 tasks (see GA):

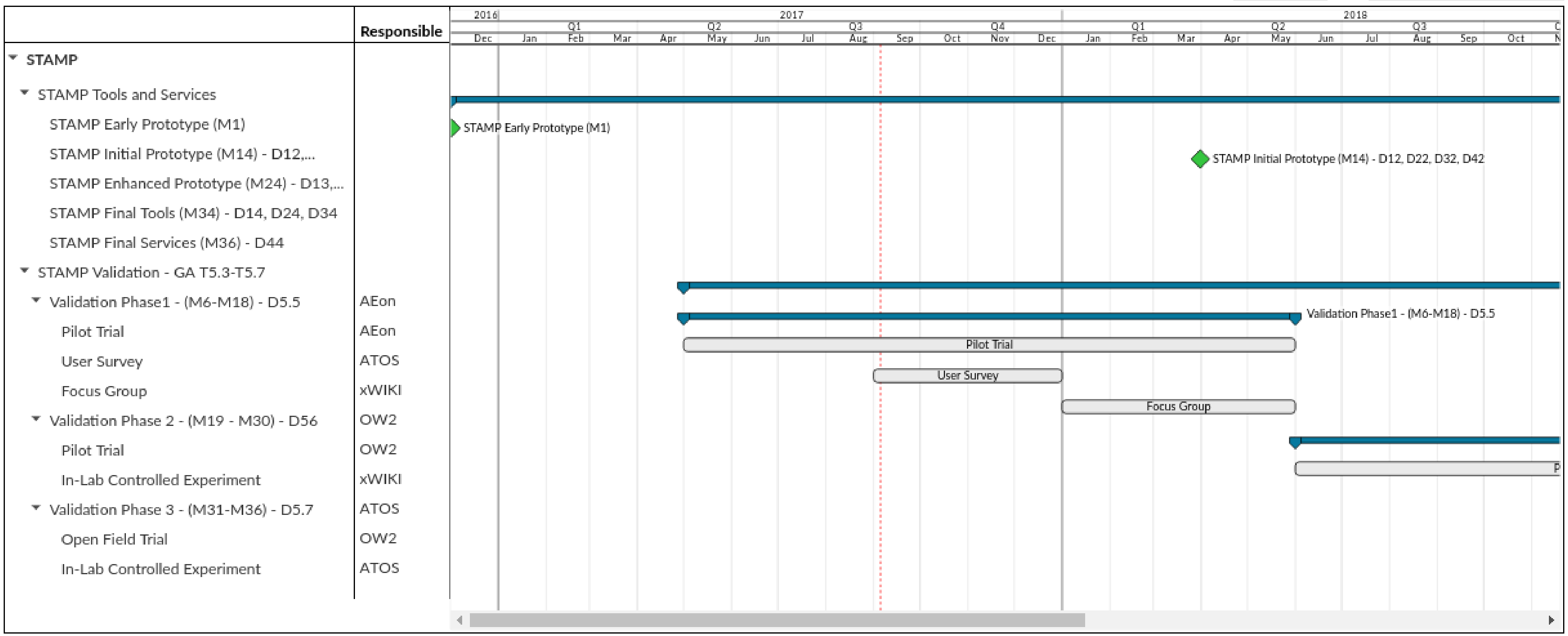
- T5.3: ProActive Scheduling and Workflows (ActiveEon) case validation;
- T5.4: FIWARE Ecosystem (ATOS) case validation;
- T5.5: TellU case validation;
- T5.6: xWIKI case validation;
- T5.7: OW2 case validation.

While in the GA the validation tasks are structured per industrial case, above activity breakdown for the roadmap is structured by validation activity type (i.e. targeting different objectives) and phases. This breakdown is more adequate for the purpose of providing a detailed roadmap than the task structure for WP5 described in the GA. There is not a direct mapping between above roadmap activities and WP5 tasks. Nonetheless, all above roadmap activities will be conducted in the context of the five STAMP industrial case tasks. Hence, they will be contributed by all the STAMP industrial partners, namely, ActiveEon, Atos, TellU, xWiki, OW2. Indeed, roadmap activities will make used of the resources available for each industrial partner in the context of the GA tasks T5.3-T5.7. In terms of periodic reporting, above evaluation activities in Table 2 will be justified in the context of GA T5.3-T5.7 as well.

Figure 1 and Figure 2 show the Gantt diagram of the evaluation roadmap[2], aligned to the STAMP release plan (shown at the top of the diagram). *Responsible* column shows the coordinator of each validation activity. Remaining industrial partners are contributing to each validation activity, as mentioned in Table 2[3].

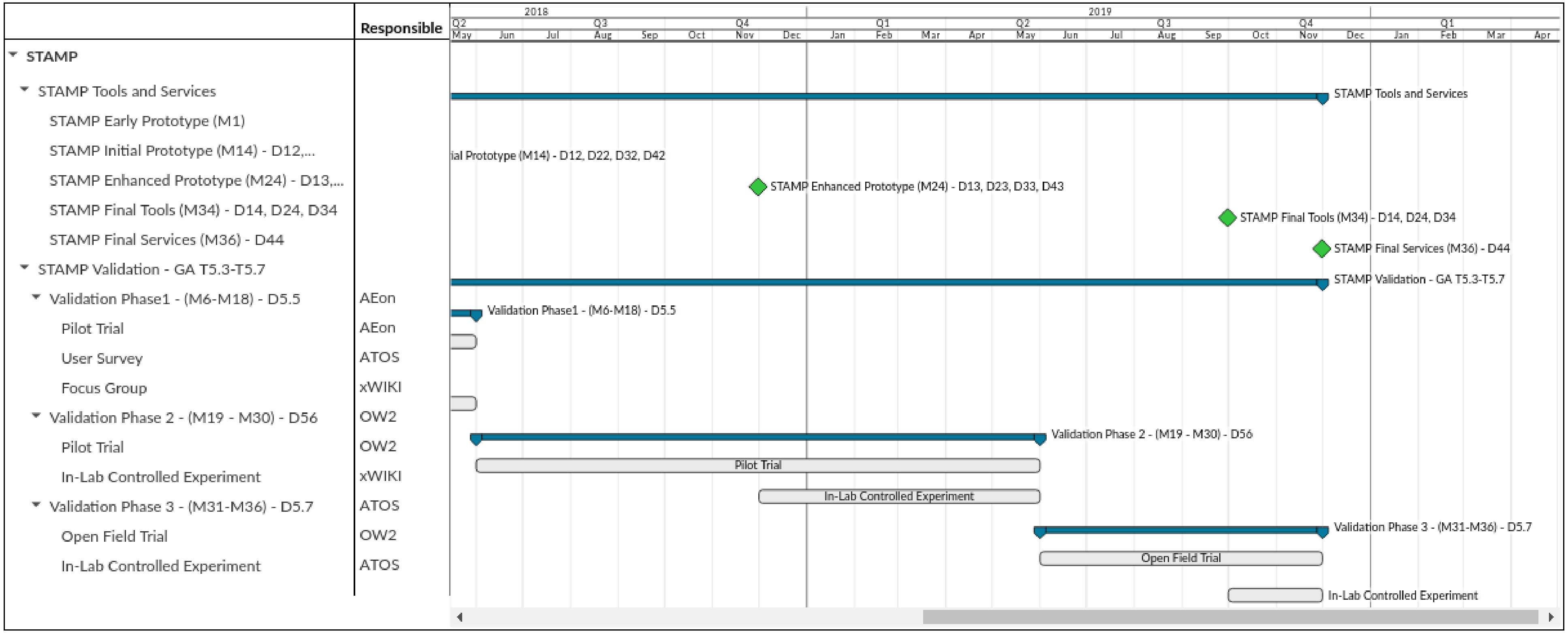
[2] Gantt diagram is split into two figures in order to improve its readability

[3] Contributors are not explicitly mentioned in the Gantt diagram in order to keep it simple.



	Responsible
STAMP	
STAMP Tools and Services	
STAMP Early Prototype (M1)	
STAMP Initial Prototype (M14) - D12,...	
STAMP Enhanced Prototype (M24) - D13,...	
STAMP Final Tools (M34) - D14, D24, D34	
STAMP Final Services (M36) - D44	
STAMP Validation - GA T5.3-T5.7	
Validation Phase1 - (M6-M18) - D5.5	AEon
Pilot Trial	AEon
User Survey	ATOS
Focus Group	xWIKI
Validation Phase 2 - (M19 - M30) - D56	OW2
Pilot Trial	OW2
In-Lab Controlled Experiment	xWIKI
Validation Phase 3 - (M31-M36) - D5.7	ATOS
Open Field Trial	OW2
In-Lab Controlled Experiment	ATOS

Figure 1 STAMP Validation Timeline (Gantt Diagram) – Phase I



	Responsible
STAMP	
STAMP Tools and Services	
STAMP Early Prototype (M1)	
STAMP Initial Prototype (M14) - D12,...	
STAMP Enhanced Prototype (M24) - D13,...	
STAMP Final Tools (M34) - D14, D24, D34	
STAMP Final Services (M36) - D44	
STAMP Validation - GA T5.3-T5.7	
Validation Phase1 - (M6-M18) - D5.5	AEon
Pilot Trial	AEon
User Survey	ATOS
Focus Group	xWIKI
Validation Phase 2 - (M19 - M30) - D56	OW2
Pilot Trial	OW2
In-Lab Controlled Experiment	xWIKI
Validation Phase 3 - (M31-M36) - D5.7	ATOS
Open Field Trial	OW2
In-Lab Controlled Experiment	ATOS

2018
2019
Q2
Q3
Q4
Q1
Q2
Q3
Q4
Q1
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
STAMP Tools and Services
ial Prototype (M14) - D12, D22, D32, D42
STAMP Enhanced Prototype (M24) - D13, D23, D33, D43
STAMP Final Tools (M34) - D14, D24, D34
STAMP Final Services (M36) - D44
STAMP Validation - GA T5.3-T5.7
Validation Phase1 - (M6-M18) - D5.5
Validation Phase 2 - (M19 - M30) - D56
Pilot Trial
In-Lab Controlled Experiment
Validation Phase 3 - (M31-M36) - D5.7
Open Field Trial
In-Lab Controlled Experiment

Figure 2 STAMP Validation Timeline (Gantt Diagram) – Phase II and III

The remaining of this section provides details about the each validation phase. See section 0 for a description of the elements of the evaluation framework mentioned in the tables. See section 11 for a description of the elements of the evaluation method mentioned in the tables.

8.2.1. Phase 1

Table 3 Phase 1 description

Phase	Phase 1
Timeline Frame	M3-M18
Validation Objectives	● Provide an early assessment on the functionality and usability of the STAMP techniques and the toolset; ● Provide an initial empirical assessment on the practical utilization of the early and initial prototypes of the STAMP toolset; ● Involve end-users in the development process of the STAMP toolset, aiming at aligning the toolset features and usage to their industrial needs.
Validation Activity Types	● User surveys; ● Focus groups; ● Pilot trials. In the period M3-M9, before the STAMP validation framework is reported in this document, we are conducting informal, not systematic (e.g. not methodological and framework supported) validation activities, consisting of using existing early STAMP prototypes of the toolset for test amplification on the different industrial use cases. Findings (e.g. metrics reported in D5.1 about test coverage), usage issues, bugs, required features, etc. are being reported through different channels, including informal discussion threads on chats or by email, and trackers (GitHub, GitLab). After M9, a systematic validation, adopting this framework and following this proposed roadmap, will be conducted.
Validation Expected Outcome	● Early feedback about toolset features and usability for developers on technical WP1-WP4: ○ Communication of additional required features and/or refinement/improvement of existing ones; ○ Reporting on toolset utilization issues (e.g. bugs, usage limitations, configuration, etc.); ● Industrial scoped recommendations for toolset improvement.
Reporting	D5.5 UC validation report V1, M18

8.2.2. *Phase 2*

Table 4 Phase 2 description

Phase	Phase 2
Timeline Frame	M18-M30
Validation Objectives	● Provide an empirical assessment on the practical utilization of the stable prototypes of the STAMP toolset; ● Involve end-users in the development process of the STAMP toolset, aiming at aligning the toolset features and usage to their industrial needs; ● Assess the benefits of adopting the STAMP toolset on industrial cases compared to those gained using third-party solutions.
Validation Activity Types	● Pilot trials; ● Comparative experiments. In both validation activities, we conduct a systematic, validation framework assisted assessment of the STAMP toolset adoption on real industrial cases, in Lab controlled testing environments.
Validation Expected Outcome	● Feedback about toolset utilization for developers on technical WP1-WP4: ○ Reporting on issues (e.g. bugs, usage limitations, configuration, integration, etc.); ● Reports on KPI assessment and hypothesis fulfillment: ○ KPI metric figures for pilot trials and comparative experiments; ● Industrial scoped recommendations for toolset improvement.
Reporting	D5.6 UC validation report V2, M30

8.2.3. Phase 3

Table 5 Phase 3 description

Phase	Phase 3
Timeline Frame	M30 - M36
Validation Objectives	● Provide an empirical assessment on the practical utilization of the final stable prototypes of the STAMP toolset; ● Assess the benefits of adopting the STAMP toolset on industrial

	cases compared to those gained using third-party solutions; ● Promote the adoption of STAMP techniques and methods in open communities of developers; ● Collect feedback from open communities of developers on their usage of the STAMP toolset in real development situations.
Validation Activity Types	● Open Field Trials; ● Comparative experiments.
Validation Expected Outcome	● KPI hypothesis assessment: ○ Compared KPI metric figures, adopting STAMP toolset and competitive solutions; ● Final assessment conclusions; ● Industrial partner recommendations for future work.
Reporting Date	D5.6 UC validation report V2, M36

9. Conclusion

This document describes the method, framework and roadmap for the industrial-driven evaluation of the STAMP toolset and services. The method defines the process to conduct the evaluation activities. The framework provides the tools for supporting the evaluation activities. The roadmap defines the timeline of the activities, in agreement with the development roadmap. These elements, working together, will drive the STAMP evaluation and the reporting of resulting findings and recommendations, aiming at improving the effectiveness, efficiency and user-satisfaction of industrial adopters on their real-life testing situations that require test amplification.

This document provides the initial specification. On M20, a refined version will be release, focusing on concretizing these elements: method, framework and roadmap, to the actual utilization adopted in the following M9-M20 period of evaluation activities, paving the way for the remaining evaluation until the end of the project.

10. Appendix A: Elements of the Evaluation Framework

In the following subsections, the concepts and constituents of the validation framework are introduced in detail.

10.1. Quality Model, Requirements and Metrics

The overall goal of validation is to make the software usable in a broad sense i.e. “the user can do what he wants to do the way he expects to be able to do it, without hindrance, hesitation, or questions” [3]. However, a broad range of characteristics can influence the quality of system, as defined in industry standards such as the ISO/IEC 25010:2011[4]. This standard is part of a suite known as SQuare (ISO/IEC System and Software product Quality Requirements and Evolution). It defines a hierarchical **quality model** [5], a model that defines a set of quality entities and their relationships, which allows the formal specification of quality requirements and their usage on the evaluation of software quality. Leaf quality entities in this hierarchical tree model are measurable, that is, they are metrics, which enable the quality assessment of software.

From the broad range, taking into account the purpose of the usage of the software product, a narrow set of aspects are chosen, influenced from ISO 25010 quality model, to drive the evaluation process. These aspects have been selected based on the typical categories of requirements related to the software that are considered in the STAMP validation process, to specify detailed objectives of user acceptance validation. These categories are as follows:

- Functionality
- Information
- Usability
- Performance
- Benefit/impact

These categories are described below in Table 6 in more detail. For each category, the objective and the description in terms of validation questions and metrics that will be considered during validation are presented.

Table 6 Quality model for validation

Functionality Assessing whether the functional capabilities of STAMP toolset and services meet the user’s needs	Examples: ● Does STAMP improve the test coverage for Java based projects? ● Does STAMP generate test cases that detect regressions? ● Does STAMP help to find optimal deployment configurations that maximizes the performance of services?

[4] ISO/IEC 25010:2011, “Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models”, 2011 http://iso25000.com/index.php/en/iso-25000-standards/iso-25010

	● Does STAMP help to generate test cases that reproduces exceptions shown in runtime logs? Metrics: ● Functional metrics defined in T5.1 (reported in D5.1 [7]) for identified KPIs; ● New/refined elicited functional requirements; ● Number of issues/bugs reported and collected in trackers related to the STAMP functionality: percentage of issues fixed; ● Qualitative subjective user's perception on the functionality, collected through questionnaires in validation activities involving external users.
Information Assessing whether STAMP toolset and services are adequately supported by accompanying information and support, so users can fully exploit their functionality	Examples: ● Do the users understand the purpose and the usage of STAMP tools? Do they miss any information? ● Do they find information support for using the tools in their daily test amplification tasks? ● Are the users blocked because they can't find the procedures to use the tools in concrete tasks? Metrics: ● Number of questions posted by users to the developers (for instance in a private (i.e. for STAMP consortium) or public (i.e. for external communities) forums/mailing lists[5]). Percentage of them correctly answered (e.g. voted as positive by users); ● Number of validation tasks uncompleted during comparative studies (over the total) caused by a wrong usage of the tools; ● Qualitative subjective user's perception on provided information collected through questionnaires in validation activities involving external users.
Usability Assessing whether the STAMP toolset and services provides an intuitive interface that facilitates the user's adoption	Examples: ● Are the toolset interfaces intuitive enough, easy to understand, learn and use by average users, depending on adopted role? ● Do the users feel comfortable using this interface or do they require changes on it?

[5] The private STAMP forum or the mailing list are hosted by OW2 during the project lifetime and maintained by STAMP developers. The public STAMP forum or the mailing list could be also hosted as an OW2 open-source project, being initially maintained by the STAMP developers and later on by the open-source communities.

	Metrics: ● Number of reported usability issues; ● Number of validation tasks uncompleted during comparative studies (over the total) caused by a wrong usage of the tools; ● Time required to complete validation tasks (computing only time required to manage the interface); ● Qualitative subjective user's perception on usability collected through questionnaires in validation activities involving external users.
Performance Assessing whether the STAMP toolset and services offers a competitive efficiency on completing the test amplification tasks	Examples: ● How does STAMP toolset and services performs in terms of efficiency and stability attending to the scale of the input source code? Metrics: ● Metrics defined in D5.1 for associated performance KPIs; ● Qualitative subjective user's perception on the performance, collected through questionnaires in validation activities involving external users.
Benefit/Impact Assessing the perceived benefit STAMP provides to users and the value of using STAMP	Examples: ● Is the STAMP toolset and services appropriate for the purpose it is designed? ● Does STAMP provide added-value to the current test phases on software engineering? ● What are the benefits users perceive from using the STAMP toolset? Metrics: ● Qualitative subjective user's overall satisfaction of using STAMP toolset, collected through questionnaires in validation activities involving external users; ● Perceived benefits of STAMP and level of impact on software development value chain reported by users; ● Efficiency and completeness of fulfilling test amplification tasks with and without STAMP (effectiveness); ● Time reduction of fulfilling tasks with and without STAMP (productivity); ● User engagement level: number of participants on public online field trials.

The assessment of above validation objectives will be verified by computing the metrics associated to each requirement type (e.g. aspect) in above classification. In general terms, these metrics can be classified into:

Qualitative metrics: these metrics are suitable to qualify subjective opinions from STAMP users. They will be measured (from individual users) using:

- Textual feedback, when the subjective opinion cannot be foreseen beforehand. Results are processed manually;
- 5-Likert gauges, when the feedback type is identified beforehand and included in the evaluation questionnaires. Collected results can be processed statistically, showing relevance depending on the number of responses.

Quantitative metrics: these metrics are suitable to quantify objective facts that characterize diverse aspects on the usage of the STAMP toolset and services on controlled experiments. They should not depend on experimenters' experience and background (e.g. unbiased results). In this group we include:

- Metrics defined in D5.1 [7] for selected KPIs;
- Productivity gains (for instance in productivity) measured in lab controlled experiments, showing the benefits of adopting STAMP in real industrial scenarios, as expressed by computed KPIs. Eventually, in those industrial scenarios that previously adopted some state of the practice techniques targeting STAMP functional aspects (e.g. better test coverage, regression bug detecting, etc.), the measured productivity gains will be also compared to these state of practice baseline;
- Number of reported issues or bugs (and the percentage that are addressed or fixed);
- Number of requested features (and the percentage of them, that having being accepted in the scope of the STAMP project, are implemented);
- Number of detected toolset usage errors during controlled experimentation, etc.

10.2. Validation Target Groups

STAMP methods and techniques, implemented as a toolset and public services, are intended to be adopted by different types of users when performing common test amplification tasks in different phases of the software development life-cycle. Therefore, one of the elements of the validation framework is the identification of these roles, among the potential adopters of the STAMP technologies, together with their needs and expectations, by surveying the widest possible number of communities involved in software development. These roles are identified by collecting the demographic information of the participants in the different validation activities. When possible they will be associated to the different test amplification tasks and supporting tools, determining, through the responses to questionnaires to what extend the tools satisfy their needs and expectations.

A number of relevant roles have been already identified by the consortium:

- Software developer: functional development of applications, service and their features;
- Test developer: development of tests cases, including unitary, integration, system tests, etc.;
- Software tester: execution and analysis of test cases; quality assurance;

- DevOps integrator: integration and deployment of software application and services. Optimal deployment configuration for selected non-functional indicators;
- Software maintainer: evolution of software application and services with new features. Maintenance of software releases.

As part of the collected demographic profile for each participant in the evaluation activities, the following information will be requested:

- Technological expertise, focusing on:
  - Programming languages;
  - Testing technologies and methodologies;
  - Software management and DevOps technologies.
- Industrial background, focusing on:
  - Academic studies;
  - Years of experience;
  - Operational domain (e.g. target market) of the company;
  - Adopted roles within the software engineering lifecycle.

Demographic profiles will be taken in consideration when preparing the validation groups (i.e. for instance adopting random sampling strategies) in order to reduce the bias possibly introduced in the validation results because of an unbalanced selection of these target groups.

10.3. Validation roles

Different roles are involved in the validation activities:

- **participants** are directly involved in the assessment activities and constitute the validation target groups (see section 10.1). They conduct the experimentation by performing proposed validation tasks on concrete target systems, adopting a treatment under assessment (e.g. the STAMP toolset);
- **facilitators** are involved in some concrete validation activity types (e.g. comparative studies, see section 10.7) organizing the validation target groups, coordinating the validation tasks and recording some metrics, but they are not part of the validation target groups (e.g. the validation subjects).

10.4. Supporting material

Validation roles are supported before and during the experimentation with some materials required to perform the validation activities. Each activity type will require different kinds of materials, which will be make available (or linked) in the STAMP website[6]:

- **Documentation**
  - STAMP project presentations (slides): introductory presentation to STAMP project, concepts, methods and tools;
  - Usage scenarios: exemplary practical real situations of STAMP methods and tools application;

[6] The private documentation for consortium usage will be restricted from public access

  - Toolset technical documentation: getting-started and user-manual documents accompanying the STAMP toolset;
  - Walk-through user guides: step-by-step tutorial explaining the usage of STAMP toolset in concrete application scenarios;
  - Demos (videos): available from STAMP website or YouTube channel, they may introduce the STAMP project in a nutshell and/or a concrete method, technique or tool;
  - Webinars: an informative online presentation of the STAMP project and/or a specific aspect. It can be supported by slides and recorded as video.

- **Questionnaires/Forms**
  - Demographic profile: metadata characterizing a concrete participant or a group of them participating in a validation activity, particularly focusing on expertise and background related to the subject of validation;
  - Feedback-collection questionnaires: online questionnaires intended to collect qualitative feedback from participants of subjective validation concerns;
  - Experiment/tasks descriptions: detailed description of the tasks (and their steps) to be conducted during some validation experiments (e.g. comparative studies);
  - Metric collection forms: intended to collect measures of quantitative metrics associated to KPIs assessed in the validation activities;
  - Requirements elicitation forms (e.g. issue trackers): STAMP provides a GitLab infrastructure that includes an issue tracker to collect functional requests. Some STAMP tools hosted on GitHub (e.g. WP1 Descartes, DSPot, etc.) use a similar issue tracker. The STAMP GitLab instance[7] is hosted by OW2 and provides support for all steps from idea to code development, code deployment and feedback gathering. As put by the GitLab team, these steps consist of the following: idea, issue, plan, code, commit, test, review, staging, production, feedback. Ideas and general discussions can take place over a chat engine (e.g. Rocket.chat, an OW2 project providing powerful chatting capabilities), issues can be submitted and discussed over the GitLab issue tracker, planning can take advantage of the GitLab issue board;
  - Issue/bug tracking tools: using the same tracker mentioned above.

- **Experimentation environment**
  - STAMP Toolset Setup: preconfigure executable (standalone) STAMP toolset virtual environment, ready for experimentation,
  - Experimental source code projects: pre-existing source code projects intended to be used for test amplification during some validation activities, such as comparative studies.

- **Reporting forms**: agreed template for reporting validation results analysis and findings.

[7] The exiting STAMP GitLab instance has restricted access to the project consortium. When released to the public, the STAMP tools releases will be hosted in the public OW2 forge.

Some materials will be authored by the validation work-package (WP5), such as usage scenarios, demos, reporting forms, etc. and others by the technical work-packages (WP1-WP4)[8], such as technical documentation, webinars, experimental environment, etc.

As a result of the validation activities, feedback will be duly reported to the STAMP toolset and services development teams[9]. There are different goals for providing this feedback:

- Provide early feedback to developers aiming at supporting them to improve the functionality, usability, efficiency, effectiveness and other aspects of the STAMP toolset and services and cover a wider application range on industrial real situations:
    - New functional (and nonfunctional) requirements and features;
    - Detected issues (and possible bugs) encountered during the utilization of the tools and services on industrial situations requiring test amplification;
    - Identified usability issues that makes difficult or hampers the utilization of the tools under the situations required by adopters;
- Report usage experiences adopting the toolset and services on concrete industrial scenarios of code testing;
- Assess the fulfillment of the hypothesis associated to the baseline KPIs defined in the GA (see [7]) and provide quantitative estimations of the benefits of adopting STAMP technologies in industrial real applications;
- Assess the efficacy and efficiency of the STAMP methods and their toolset implementation (eventually compared to the equivalent results formerly obtained by adopting state of the practice techniques, when they were available for some industrial cases).

10.5. Reporting

Reports will be structured according to common templates depending on the type of reporting findings. These templates describe the required information and the format adopted to structure it:

- Requirements (new, amendments) will be reported according to the standard issue schema adopted by the STAMP project issue tracker tools (e.g. GitLab, Github) to report features (epics or user stories) [8]. See table below for feature issues;
- Results of comparative experiments with assessment of KPI hypothesis (see section 10.7.2.3) will be reported adopting proposed standards of scientific reporting guidelines for controlled experiments in software engineering [6]. Details of the adopted reporting guidelines will be included in the next version of this deliverable (e.g. D5.4);
- Issues/bugs will be reported according to an agreed issue schema adopted by the STAMP project (see Table 7) issue tracker tools to report issues or bugs, based on common tracker schemas such as Jira, Bugzilla, MantisBT, etc[10]:

[8] WP1-WP3 will contribute to the technical aspects of the documentation, while WP4 will contribute to the usability aspects, including users' manuals for available tool chains and tool clients, etc.

[9] WP1-WP3 development teams will take care of the feedback related to the core functionality and implementation of the STAMP tools, while WP4 will take care of the feedback related to the tool clients, services, interoperability aspects, usability, etc.

[10] https://marker.io/blog/bug-report-template/

Table 7 Template for issue reporting: new features and bugs

Issue <Number>	Issue Id, provided by the tracker
Key	A brief one-line summary of the issue
Type	Type of reported issue. It could be: ● **Bug**: issue perceived by the reporter as a potential bug, which has to be confirmed by assignee; ● **Feature**: issue describing a new requested functionality or a non-functional property to be supported.
Tags	Above issue type can be further refined, in the case of bugs, by adopting a number of predefined tags (taken from a proposed STAMP tag cloud), including[11] REGRESSION, CONFIGURATION, PERFORMANCE
Description	A detailed description of the issue. For features, this section should provide a functional description of the required functionality. When describing features formally as user stories, the description can include this formal syntax: *As a <**role**>, I can <**activity**> so that <**business value**>* With this form, all the stakeholders involved in the requirement analysis can understand both the role of the user and the business benefit that the new functionality provides. For bugs, this section should describe as well: ● the observed execution behavior and obtained results; ● the expected execution behavior and results.
Reproducibility *<For bugs only>*	The occurrence of the issue, that is, the likelihood to be reproduced. Possible values: always, sometimes, random
Severity *<For bugs only>*	The degree of impact (as perceived by the reporter) the bug has on the operation of the tool/service being used in a concrete industrial situation. Possible values are: critical, major, minor, trivial
Tool/Service/Component Version	A reference name of the tool/service/component and the version the reported issue is related to (or affecting to)
Execution Environment **Steps to reproduce** **Snapshots**	A detailed description, step-by-step of the procedure followed by the reporter to reproduce the bug reported. This section should also include a description of the execution environment (platform,

[11] It is not an exhaustive tag list at the time of releasing this document.

Other files and URLs *<For bugs only>*	OS, etc.), including information about the version of the executed STAMP tools/services and their local dependencies (in case of standalone execution). Additional visual proofs, such as snapshots, providing additional visual information of the bug can be included, as well as input files required for reproducing the bug or URLs pointed to the sources of such inputs.
Relationships	A list of relationships to other issues. In case of features, these relationships can be used to structure them, grouping related features. Possible relationships: ● Child of / Parent of ● Related to ● Depends on
Reporter: **Name, email**	Reference information of the reporter, so the assignee can contact back for further issue refinement, if needed. A reference to the reporter could be automatically added by the tracker.

- KPI metrics will be reported according to the information schema adopted in T5.1 and reported in D5.1 [7]
- Qualitative (and subjective) findings, based on user feedback collected in different validation activities will be reported by adopting the information schema depicted in the following Table 8

Table 8 Reporting schema for validation findings

Finding <Number>	Finding descriptive title
Rationale	Reasoned description of the finding
Aspect	List of categories (see section 10.1) the finding belongs to
Related to	List of STAMP toolset components, services or functional aspects (e.g. unit test amplification, etc.)
Priority	Importance given to the finding by stakeholders or evaluators
Reported on	Evaluation activity (phase, type) from where the finding was inferred.

The project adopts an iterative and frequent reporting style, fostering an agile communication between UC evaluators (in WP5) and technical developers in WP1-WP4, who provide first[12] (WP4) and second[13] (WP1-WP3) level support, through:

- On continuous basis internal reporting: duly providing feedback, typically toolset usage experiences, bug/issues or requirements as a result of continuous evaluation (i.e. focus groups, pilot trials) by UC partners involved in the project, using direct or indirect internal communication channels (i.e. email, chat, wiki, tracker, etc.);
- Discrete official reporting: at fixed milestones, through official WP5 deliverables (D5.5-D5.7). These reports will include detailed reporting of the validation activities conducted during the validation period (including both continuous and discrete activities), description of collected data, their analysis and results.

10.6. Validation Environments

The validation activities are conducted under different experimentation environments, characterized by their context and their available control mechanisms to drive the assessment activities, namely:

Lab controlled environment

In a Lab controlled environment there is a precise management of the different factors or conditions that can influence the execution of the assessment activities and their results. All these factors are controlled by the validation facilitators, who take care of preparing the assessment setup to concrete values. In particular, a controlled assessment environment includes, among others, some of these features:

- Concrete evaluation goals and associated verifiable hypotheses
- Specific setup of control and treatment methods and techniques, which constitutes the toolset target of the assessment
- Proper selection of experimenters' groups, suitably chosen to reduce assessment bias
- Similarly, a proper experimentation plan, which combines experimentation tasks and conducting groups also aiming to reduce bias on the assessment findings
- Specific measurement frameworks, collecting different metrics during the execution of the experimentation tasks
- Observed experimentation (by facilitators) who take care of record additional experimentation metrics
- Formal analysis of experimentation results
- Formal reporting of validation findings

Lab controlled validation activities can be conducted on continuous basis, but typically are also conducted in discrete events, such as workshops.

[12] First level support on tool client usage
[13] Second level support on tool internal operation

Online field environment

Unlike Lab controlled environment, online field environment shows little control over the experimentation context under which groups of evaluators are conducting assessment experiments, typically because:

- Assessment activities are conducted on the real-world context of the evaluators themselves, which are unknown and/or out of control for facilitators of the validation
- The evaluation activities are conducted by open communities of potential adopters, having neither control of the participants and its number nor of their demographic profile

The precise experimentation conditions are not held in the online field environment.

However, both approaches are complementary. Under lab controlled environment, evaluation hypothesis can be accurately verified scientifically, resulting in concluding findings. However, these concluding facts are only true under very concrete experimental conditions that rarely resemble practical situations. The online field environment enables the assessment of STAMP toolset under real and practical situations, not constrained to predefined conditions, and assessed by a larger group of practitioners. However, the accuracy of findings cannot be determined unless by the large statistical significance of the sampling.

10.7. Validation Activity Types

A number of different validation activities have been identified and designed with the purpose of validating the STAMP toolset and services addressing different validation goals. These types can be roughly classified according to previous taxonomy of validation environments into online field and lab controlled. See section 11 for a description of the common tasks mentioned in the validation method row.

10.7.1. Online/Field

10.7.1.1. User surveys

Table 9 User Survey description

Validation Activity Type	User Survey
Timeline Frame	**Phase**: Phase 1 **Event type**: Discrete Event
Target User Group	**Participant type**: ● UC partners' members not involved in the project (2-4 participants per partner); ● Members of external organizations; ● No previous knowledge about STAMP project, aiming at reducing bias on results. **Number of participants**: 10-20 participants

Activity Goal	Collect participants' feedback about the STAMP approach, concept, toolset functionality, and the baseline of KPIs proposed in the GA. Collected feedback can be structured as toolset new requirements or refinements of existing ones.
Supporting materials	Participants are introduced on the STAMP project objectives and toolset features, using different materials: ● Slide-based presentation; ● Usage scenarios: they describe the adoption of STAMP concepts, methods and techniques in concrete and real situations, which exemplifies the practical usage of the STAMP toolset for potential adopters; ● "STAMP for dummies" manuals.
Validation method	1. Establish the goal of evaluation; 2. Identify tool and services target of evaluation; 3. Plan evaluation activities; Dedicated session (1-2 hours) attended by external participants and STAMP consortium representatives: • External participants are introduced to the STAMP project concepts, methods, toolset and usage scenarios (1 hour); • External participants are prompted to provide feedback about the project (1 hour or less), including: o Demographic information; o Feedback about the STAMP project from industrial adoption point of view, by answering an online (i.e. Google Forms) questionnaire that includes: ▪ General questions about the project; ▪ Usage scenario specific questions; • In the online questionnaire, participants can also express new functional requirements and KPIs, amend or refine the functionality already described for the STAMP toolset and/or rank them; • The questionnaire can capture subjective responses with 5-Likert gauges and text boxes for comments; • Assess the evaluation results; • Author the evaluation report;

10.7.1.2. Open field trial

Table 10 Open field trial description

Validation Activity Type	Open Field Trial
Timeline Frame	**Phase**: Phase 3 **Event type**: Continuous Event
Target User Group	**Participant type**: Developers of open-source communities, such as OW2, FIWARE or Eclipse communities, and also the academic community. **Number of participants**: Unbound
Activity Goal	Evaluate the STAMP concepts, methods and toolset in daily practical testing situations dealt with by developers engaged in open-source development communities
Supporting materials	Participants engaged in this validation activity are supported by different materials publicly available online in the STAMP website: ● Stable STAMP toolset and services: ○ Downloadable standalone toolset, for supported platforms; ○ Online accessible STAMP services; ● Public documentation (with links accessible from STAMP website): ○ Webinars, demos, video channel in Youtube, presentations in Slideshare, public repository readme s and getting started documents; ○ Scientific articles describing the underpinnings of each tool; ● Means to collect feedback from STAMP adopters: ○ Online questionnaires (i.e. Google Forms) ■ Links to these questionnaires will be advertised in STAMP website, public documentation, mailings, newsletters, etc.; ○ A public mailing-list letting the participants to get support in their process of adopting the STAMP toolset and services; ○ Feedback forms embedded in the STAMP toolset and/or public services; ○ Public toolset tracker in OW2 GitLab or Github; ○ Web analytics (i.e. Google Analytics).
Validation method	There are several features that characterize the validation method adopted by this validation activity type. The tasks for validation method described in section **Erreur ! Source du renvoi introuvable.** are not applicable here, as the validation objectives are determined by the

	participants themselves, outside of the STAMP consortium control: ● Free experimentation: participants are free to use the STAMP toolset and services as they need for different personal purposes, according to their own circumstances and adopting their own experimentation method, not being possible nor desirable to impose a concrete one from the STAMP project; ● Toolset and services usages is guided by public documentation; ● Usage feedback is collected through online questionnaires and feedback forms; ● Other possible feedbacks are collected by the toolset trackers, as reported issues or bugs, as well as requests for additional functionalities; ● Metrics will be reported accounting for qualitative findings and qualitative number of reported issues.

10.7.2. Lab controlled

10.7.2.1. Focus groups

Table 11 Focus group description

Validation Activity Type	Focus Groups
Timeline Frame	**Phase**: Phase 1 **Event type**: Continuous or discrete (e.g. workshop based)
Target User Group	**Participant type**: ● UC partners' members involved and external to the project (2-3 participants per partner); ● Participants must adopt (in real situations) roles that match the STAMP target roles. **Number of participants**: 10 - 15
Activity Goal	The refinement, formalization and prioritization of STAMP requirements and KPI: ● Existing requirements and KPI defined in the GA; ● New requirements and KPI collected during the user survey.
Supporting materials	● Usage scenarios provided as supporting material for use survey ● Requirements and KPI collected during the user survey
Validation method	1. Establish the goal of evaluation; 2. Identify tool and services target of evaluation 3. Plan evaluation activities:

	● Collect current industrial testing practices without adopting STAMP methods and techniques; ● Introduce STAMP concepts, methods, techniques and scenarios of usage; ● Systematic requirement and KPI refinement, prioritization and association of metrics to KPIs: ○ Discussion will be recorded for further reference and analysis. 4. Refine the quality model for evaluation; 5. Assess the evaluation results; 6. Author the evaluation report.

10.7.2.2. Pilot trials

Table 12 Pilot trial description

Validation Activity Type	Pilot Trial
Timeline Frame	**Phase**: Phases 1 and 2 **Event type**: Continuous event
Target User Group	**Participant type**: ● Members of the industrial UC partners; ● Active member participation in STAMP project, hands on the toolset. **Number of participants**: 10 - 15
Activity Goal	● Reporting usage experiences of toolset/service application on concrete industrial scenarios of code testing; ● Reporting issues and potential bugs experienced during toolset/service utilization; ● Assessment of concrete requirements and KPI, collecting and reporting associated metrics.
Supporting materials	● Available prototypes of STAMP toolset and services; ● Toolset/service developers support (e.g. installation, configuration, usage); ● Toolset/service internal documentation: wiki; ● Framework for metric measurement and reporting; ● Framework for feedback reporting: tracker, chat, email.
Validation method	1. Establish the goal of evaluation; 2. Identify tool and services target of evaluation; 3. Refine/Specify the quality model (optional);

	4. Specify quality requirements; 5. Select the quality metrics; 6. Define criteria for assessment; 7. Plan evaluation activities: ● Define common test amplification tasks, based on aforementioned usage scenarios; ● Specialize or customize above tasks for each industrial UC; ● Define additional UC specific amplification tasks if needed; ● Continuous-base experimentation, adopting STAMP methods and techniques, using STAMP toolset and services, performing common and specific above tasks; 8. Assess the evaluation results; 9. Author the evaluation report: ● Report experimentation findings and issues, through identified communication channels back to WP1-WP3 developers: ○ Daily usage reporting; ○ Reporting on official deliverables.

10.7.2.3. In-Lab controlled experiments

Table 13 In-Lab Controlled experiment description

Validation Activity Type	In-Lab Controlled Experiment
Timeline Frame	**Phase**: Phases 2 and 3 **Event type**: Discrete (e.g. workshop) event
Target User Group	**Participant type**: ● Industrial UC partners; ● No previous experience in STAMP project; ● Background on software development and testing, adopting roles targeted by STAMP toolset. **Number of participants**: 15 - 30
Activity Goal	Assess the efficacy and efficiency of the STAMP methods and their toolset implementation in test amplification tasks conducted in common industrial scenarios, confirming the hypothesis associated to the baseline of KPIs proposed in the GA. Eventually, comparative analysis of KPI results can be conducted for those industrial cases that previously adopted some of the state of practice techniques associated to the

	STAMP functional objectives.
Supporting materials	Participants engaged in this validation activity are supported by different materials publicly available online in the STAMP website: ● Stable STAMP toolset and services: ○ Pre-built, pre-configured, toolset experimentation environment, packaged for being unzipped or directly executed (e.g. a Virtual Machine, Docker container): ■ It includes all artifacts required to conduct the experimentation tasks, such as source code and development projects and other required tool dependencies; ○ Online accessible STAMP services; ● Training documentation (with links accessible from STAMP website or included within the package experimentation environment): ○ They describe the usage of STAMP toolset and services; ● Documentation describing the experimentation tasks to be conducted during the workshop; ● A metric measurement framework, also included within the package experimentation environment; ● Online questionnaire or forms for reporting experimentation results.
Validation method	1. Establish the goal of evaluation; 2. Identify tool and services target of evaluation; 3. Refine/Specify the quality model (optional); 4. Specify quality requirements; 5. Select the quality metrics; 6. Define criteria for assessment; 7. Plan evaluation activities: The following bullets depict the validation method adopting in comparative experiments. New release of this deliverable will further formalize this method. ✚ Two groups of experimenters (G1, G2) will be formed, randomly chosen to avoid any bias on the selection of groups based on the participants' expertise and background; ✚ Both groups will be trained on the purpose of the experiments and the technologies under scrutiny: ○ Introduction to testing and test amplification; ○ Training on experimentation treatment: ■ Introduction to STAMP test amplification tools: ✚ Concrete experiments will be designed, combining the groups of experimenters with a concrete combination of

	treatment and control tools for conducting the experimentation tasks: ○ In treatment tasks, STAMP toolset will be used; ○ In control tasks, current state of practice techniques for solving the proposed task will be used[14]; ✚ A set of common test amplification tasks will be proposed: ○ These tasks will be customized to different UCs showcased during the experimentation; ○ Proposed tasks are simple enough (in terms of complexity) as to be successfully completed by experimenters within a reasonable timeframe, and similarly affordable by adopting both the treatment and the control toolset; ✚ Following the experiment design, both groups (G1, G2) will perform these tasks adopting the treatment and the control toolset; ✚ During experimentation, experiment facilitators (e.g. members of STAMP consortium) will record the experiments, collecting different data, manually or assisted by a measurement framework: ○ Number of complete tasks: effectiveness; ○ Issue preventing the completion of tasks; ○ Issues faced by experimenters during the task execution; ○ Task completion time: productivity, efficiency; ○ For selected KPIs, control and treatment metrics; ✚ After experimentation, experiment facilitators will request participants to provide post-experimentation feedback, by filling in a questionnaire; 8. Assess the evaluation results: ✚ Analysis of collected data; ✚ Assessing experimental hypothesis; 9. Author the evaluation report: ✚ Reporting in deliverables: ○ KPI metrics measurements; ○ Participant feedback; ○ Hypothesis assessment; ○ Key findings; ○ Discussion about results.

[14] It only applies to those validation tasks conducted in the context of an industrial case where an alternative state of practice was previously adopted. If there are not such practices, the treatment tasks will not be conducted, and no comparative analysis of hypothesis will be performed.

11. Appendix B: Evaluation Method

The evaluation of the STAMP toolset and services show multiple facets: it involves different target stakeholders (see section 10.1), it pursues different evaluation objectives (see section 7.1) and it conducts several evaluation activities (see section 10.7). Therefore, adopting a single evaluation method, by customizing one or several existing standards, could not be possible. On the contrary, it may require the specification of tailored methods for each validation activity, according to the different pursued validation objectives. The approach adopted here consists in collecting a set of standard validation tasks that suit the different validation needs. For each validation activity, a number of selected tasks are aggregated to create a tailored validation method. These tasks constitute process fragments that, when tiled together, constitute a complete customized process. This approach based on process fragments has been adopted, for instance, in the customization of business processes [9]. In the following, the initial set of validation tasks adopted in the validation activities outlined in section 10.7 is described. These validation tasks are inspired by existing standards, such as the ISO/IEC 25040:2011[15], or proposed reporting guidelines for controlled experiments in software engineering [6].

11.1. Establish the goal of evaluation

The purpose of the evaluation should be established in clear terms by defining the evaluation goals. They should be aligned to the overall objectives of the STAMP project as they were stated in the GA. These goals are identified and agreed by the stakeholders involved in the evaluation activities, including industrial adopters (either within the consortium UC partners or within external members of the open communities of developers) and members of the STAMP toolset development teams. Concretely, evaluation goals are determined within each evaluation activity (see section 10.7) at the beginning of the activity. The stakeholders that agree on the evaluation goals are WP5 industrial evaluators, WP1-WP3 technical providers and WP4 industrialization providers.

11.2. Identify tools and services target of evaluation

The specific STAMP toolset parts or services that will be the target object of a concrete evaluation activity should be identified among the complete set of STAMP tools available at that moment, together with the specification of the concrete release that will be evaluated. At the time of releasing this document, there are a number of identified STAMP tools, namely: Descartes PITest, DSpot and XWiki tool (WP1), Configuration Test Amplification Tool (WP2), Log Analyzer Tool (WP3).

[15] ISO/IEC 25040:2011, “Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Evaluation process”, 2011: https://www.iso.org/standard/35765.html

11.3. Refine/Specify the quality model for evaluation

Depending on the goals of the validation activities and the maturity level of the STAMP toolset and services (target object of the evaluation) and the scope and context of the validation activity in specific industrial use cases or scenarios, the quality model predefined in section 10.1 may require some refinement. It can be needed to refine existing quality attributes, remove some or add others, in order to accommodate the adopted quality model. For example, depending on the toolset maturity, quality attributes associated to functional aspects, usability, performance, etc., should be reformulated in the quality model as the toolset could not be mature enough as to use strict formulations of these attributes. Finally, the concrete quality model instance adopted in each evaluation activity should be specified.

11.4. Specify the quality requirements

Once the quality model has been selected, a concrete set of leaf quality attributes or measureable requirements are selected from the model, to be used during the evaluation activity as the focused quality target. Additionally other external quality requirements (e.g. not belonging to the quality model but collected from the stakeholders in the set of elicited requirements) can be included. Different factors influence this selection, including the evaluation goals, evaluated toolset parts, specific industrial use case stakeholder's requirements, etc. For instance, each UC partner can select a concrete set of relevant KPIs as the quality model target to be evaluated in their UC during an evaluation activity.

11.5. Select the quality metrics

The restricted set of quality attributes, selected in the previous step, should be measurable, that is, they have associated metrics. In this step, concrete metrics for each attribute (or requirement) are selected, together with their bound measurement function. This mathematical function provides a quantitative value for the metrics, calculated from collected input data. Additionally, a measurement tool can be chosen from those associated to each metric to collect the input data.

11.6. Define criteria for assessment

The assessment of the quality attributes or requirements, using the measurable metrics, requires the specification of a concrete criteria, for each metric, based on rating scales consisting of ranges defined by thresholds, which represents quality levels. For example, KPIs defined in the GA establish binary scales, defined by a single threshold, so the fulfillment of the KPI can be determined by the achievement of metric measurement over such a threshold. Nonetheless, multi-range scales can also be adopted, such as the Quper model for product management decision [10].

11.7. Plan evaluation activities

This plan specifies the timeline of sub-activities (which also defines) to carry out during the validation activity. For instance, in some evaluation activity types, the quality is assessed by observing representative users carrying out representative task in a realistic context of use. For these kind of activities, such as comparative experiments, a plan of evaluation sub-activities includes (see section 10.7):

- Design experimentation task and identify context of use;
- Select evaluation participants;
- Specify data collection methods;
- Define evaluation timeline and milestones

11.8. Assess the evaluation results

Collected data is processed (e.g. using statistical mathematical methods), analyzed and compared against the criteria defined for quality assessment of each quality attribute (and metric) included as target quality model. Results are interpreted by the stakeholders according to the industrial scope and context of evaluation, constituting an overall assessment.

11.9. Author the evaluation report

The evaluation results are formally reported and notified to the interested parties, through established communication means (i.e. published documentation) and channels (i.e. WIKIs, forums, trackers, etc.), with the purpose of contributing to the fulfillment of the ultimate objectives of the evaluation. For instance, feedback such as recommendations, lessons learnt, findings, etc. on toolset usage can be communicated back to the STAMP development teams aiming to help them to improve, in next releases, those quality aspects analyzed during the evaluation activities.