Evaluation in Pilot-Based Research Projects: A Hierarchical Approach

Thomas Zefferer(✉)

Secure Information Technology Center Austria (A-SIT), 8010 Graz, Austria
thomas.zefferer@a-sit.at
https://www.a-sit.at

Abstract. The evaluation of project results is a crucial part of most research projects in the information-technology domain. In particular, this applies to projects that develop solutions for the public sector and test them by means of pilot applications. In these projects, involved stakeholders require meaningful evaluation results to derive lessons learned and to steer future research activities in the right directions. Today, most projects define and apply their own evaluation schemes. This yields evaluation results that are difficult to compare between projects. Sometimes, inconsistent evaluation schemes are even applied within a project. In these cases, even evaluation results of a single project lack comparability. In absence of coherent evaluation schemes, lessons learned from conducted evaluation processes cannot be aggregated to a coherent holistic picture and the overall gain of research projects remains limited. To address this issue, we propose and introduce an evaluation scheme for arbitrary pilot-based research projects targeting the public sector. The proposed scheme follows a hierarchical approach. Concretely, it organizes evaluation criteria on different layers of abstraction. Furthermore, the proposed scheme describes in detail necessary process steps to carry out evaluations using the defined criteria. This way, the proposed scheme enables in-depth evaluations of research projects and their pilot applications. At the same time, it assures that obtained evaluation results remain comparable anytime. The hierarchical evaluation scheme introduced in this article has been successfully applied to the international research project FutureTrust, demonstrating its practical applicability, and showing its potential also for future research projects.

Keywords: Evaluation scheme · Project evaluation · Hierarchical evaluation model

1 Introduction

During the past decades, advances in information technology (IT) have disruptively changed various parts of everyday life. While the impacts on the corporate and private sector are rather obvious, also the public-sector is increasingly

M. J. Escalona et al. (Eds.): WEBIST 2018, LNBIP 372, pp. 125–146, 2019.
https://doi.org/10.1007/978-3-030-35330-8_7

employing information technologies to save costs and to provide citizens a better service. Today, leveraging the use of information technology is on top of the agenda of most public-sector institutions. While this is a global phenomenon, especially the European Union (EU) has a long tradition in pushing the use of information and communication technologies in public administrations. Accordingly, the EU invests considerable financial resources in bringing forward its digital agenda on pan-European level [5]. One approach followed by the EU to achieve these goals is the funding of international research projects that develop innovative IT solutions for the public sector. The Large Scale Pilots (LSPs) STORK 2.0[1], PEPPOL[2], epSOS[3], or e-SENS[4] are just a few examples of recent research activities funded by the EU in this domain.

Many research activities funded by the EU follow a pilot-based approach. Accordingly, these activities develop new and innovative IT solutions and integrate them into one or more pilot applications. These pilot applications typically serve a concrete use case and satisfy a certain demand. Thus, pilot applications evaluate the applicability and the usefulness of the developed solution in practice. For instance, the EU-funded LSPs STORK and STORK 2.0 developed an interoperability layer for national electronic identity (eID) solutions [12]. This interoperability layer aimed to achieve an identity federation between different national eID systems, such as the Austrian Citizen Card [11], the Belgian BELPIC card [2], or the Estonian eID card [13]. In addition to the interoperability layer itself, STORK and STORK 2.0 have developed a series of pilot applications. They have all relied on the interoperability layer developed, and have enabled EU citizens to use their respective national eID to authenticate at electronic services provided by other EU member states. Details of these pilots have for instance been discussed by Knall et al. [10] and Tauber et al. [16].

The above example illustrates that the piloting phase constitutes an integral part of pilot-based research projects and activities. However, the piloting phase is usually limited to the respective research project's lifetime. As a consequence, even successful pilot applications typically need to be terminated at the end of a research project. Turning the pilot application into a productive and viable service is usually out of the project's scope. It is thus essential that all stakeholders involved in the project derive as many lessons learned from the piloting phase as possible. Involved stakeholders include researchers, funding bodies, and, of course, end users. After completion of a project, lessons learned from its piloting phase are crucial to turn a promising pilot into a successful productive service.

Both, researchers and funding bodies are well-aware of this fact. Thus, a sound evaluation of planned pilot applications is part of most proposals for pilot-based projects. In theory, a sound pilot evaluation appears to be an adequate method for the derivation of findings and lessons learned. Still, the current situation is often unsatisfying in practice. It can be observed that pilot

[1] https://www.eid-stork2.eu/.
[2] https://peppol.eu/.
[3] http://www.epsos.eu/.
[4] https://www.esens.eu/.

evaluations are often heterogeneous with regard to the concrete approaches followed and methods applied. This leads to a situation, in which evaluation results obtained are heterogeneous too. This heterogeneity can be observed within projects, i.e. between pilot applications, and also on a higher level between different projects. Ultimately, this leads to a situation, in which obtained evaluation results are hardly comparable. Consequently, lessons learned from conducted pilot evaluations cannot be aggregated to a coherent holistic picture. This, in turn, makes it difficult for responsible stakeholders to draw the correct conclusions from obtained evaluation results.

To address this issue, we propose a common evaluation scheme for pilot-based research projects targeting the public sector. The proposed evaluation scheme is project and pilot independent. Accordingly, it can be applied to a broad range of research projects and activities. The evaluation scheme introduced in this article relies on a hierarchical evaluation-criteria model and defines a common procedure to apply criteria based on this model. By providing a common basis for pilot evaluation, the proposed scheme ensures that obtained evaluation results are homogeneous enough to enable comparisons between different pilots. In this article, we introduce the proposed evaluation scheme in detail. Furthermore, we demonstrate its practical applicability by applying the scheme to pilot applications of the EU-funded research project FutureTrust[5].

This article is based on a paper presented at the 14th International Conference on Web Information Systems and Technologies [17]. Overall, this article is structured as follows. Section 2 defines in more detail the problem tackled by this work and briefly surveys related work. In Sect. 3, we introduce a simple project model, which also serves as basis for the proposed evaluation scheme. The evaluation scheme itself is then introduced and described in detail in Sect. 4. A first application of the proposed scheme is discussed in Sect. 5. Lessons learned from this application are sketched in Sect. 6. Finally, conclusions are drawn and an outlook to future work is provided in Sect. 7.

2 Problem Statement and Related Work

During the past years, the EU has invested significant financial resources to fund a series of research projects with the goal to improve IT services that are related to the public sector. A major driver behind these funding activities has been the aim to support the EU's strategy of a Digital Single Market [7]. According to this strategy, many funded research projects have focused on achieving cross-border interoperability between public-sector IT services of different EU Member States. Examples are the Large Scale Pilots STORK and STORK 2.0[6] focusing on eID interoperability, epSOS[7] targeting the e-health sector, or PEPPOL[8], which put an emphasis on e-procurement. Results yielded by all these LSPs have

[5] https://cordis.europa.eu/project/rcn/202698/factsheet/en.
[6] https://www.eid-stork2.eu/.
[7] http://www.epsos.eu/.
[8] https://peppol.eu/.

been consolidated by the research project e-SENS[9] . Leveraging the use of IT in public-sector use cases has also been the main goal of the international research project SUNFISH[10], funded under the EU's Horizon 2020 research and innovation programme, and of the project FutureID[11] funded under the ICT theme of the Cooperation Programme of the 7th Framework Programme of the European Commission. Scientific contributions of these projects have been discussed by Suzic et al. [15] and Rath et al. [14], respectively. This brief and by far not exhaustive overview shows that various research projects have been done during the past years aiming to bring forward the use of information technology in cross-border public-sector solutions.

Examining in more detail the surveyed projects' internal structure reveals that most of them follow a similar approach: solutions developed in the project are usually tested by means of one or more pilot applications. Furthermore, all projects foresee some sort of evaluation, where obtained results are assessed against defined criteria. However, the approaches followed by the various projects to carry out evaluations differ considerably from each other. In the worst cases, there is even no consistent evaluation method applied within a project, e.g. to evaluate different pilot applications of the project. Instead, each pilot application defines and applies its own evaluation method. This heterogeneity in applied evaluation methods has an impact on obtained evaluation results. While these results might be sufficient within the scope of a single pilot application, they cannot be assembled to a coherent big picture. This, in turn, renders the conclusive derivation of findings from available evaluation results difficult.

Of course, there are sometimes good reasons to divert from a common evaluation method and to rely on pilot-specific evaluation schemes instead. One reason can be the fact that pilot applications can undergo major redesigns during project lifetime. This can necessitate an adaptation of the planned pilot evaluation as well. Another reason to abandon a common evaluation method for all pilot applications within a project can be a high degree of heterogeneity between the pilot applications themselves, which makes it difficult to find an evaluation method that perfectly fits all pilots. Despite these valid reasons, incomparable evaluation results constitute a serious issue that threatens to decrease the overall gain of a research project for involved stakeholders.

This issue has also been recognized by the scientific community. Accordingly, interesting works exist that focus on the evaluation of research projects. For instance, Khan et al. [9] discuss the problem of evaluating a collaborative IT-based research and development project. While this work identifies relevant challenges to overcome, the proposed evaluation method has been mainly tailored to one specific project. This renders an application of the proposed method to arbitrary projects difficult. More generic evaluation methods have been introduced by Eilat et al. [4] and by Asosheh et al. [1]. Both proposals make use of the balanced scorecard (BSC) approach and data envelopment analysis (DEA).

[9] https://www.esens.eu/.
[10] http://www.sunfishproject.eu/.
[11] http://www.futureid.eu/.

However, they do not take into account the specifics of the type of projects targeted in this article, i.e. pilot-based projects aiming to improve public-sector solutions.

In summary, evaluation schemes proposed in literature usually focus on very specific types of projects or are even tailored to one single project. An evaluation scheme that can be applied to a broad range of pilot-based research projects from the public sector is currently missing. The evaluation scheme proposed in this article closes this gap.

3 Project Model

The basic aim of the work presented in this article is the development of a common evaluation scheme for pilot-based research projects. The main challenge in developing such an evaluation scheme, which is applicable to a broad set of research projects, lies in the trade-off between assuring general applicability and obtaining meaningful evaluation results. On the one hand, an abstract scheme enables a broad applicability to arbitrary research projects. However, an abstract evaluation scheme typically yields rather abstract evaluation results too, which complicates the derivation of concrete conclusions. On the other hand, a more specific evaluation scheme can consider peculiarities of a given research project or its pilot applications. However, such a specific scheme can usually only be applied to a limited set of projects or pilots. The demand for general applicability is not satisfied in this case.

To overcome this trade-off, we have based the proposed evaluation scheme on a well-defined project model. The proposed scheme can be applied to any research project complying with this model. The project model has been based on the following three assumptions, which define the scope of targeted project types:

- The proposed evaluation scheme targets pilots and projects that provide solutions for the public sector. Accordingly, the evaluation scheme assumes that an IT agenda is in place, from which goals for different projects under this agenda are derived.
- The proposed scheme targets pilot-based projects, which apply and test solutions developed in the project by means of one or more pilot applications.
- The proposed evaluation scheme assumes the research project to be carried out in multiple consecutive project phases, each comprising an own phase-specific pilot evaluation.

From these assumptions, a general project model can be derived, which serves as basis for the proposed evaluation scheme. This project model defines the general project structure as well as general project-execution and evaluation phases. The two aspects of the derived project model are detailed in the following subsections.

3.1 Project Structure

From the assumptions made above, the general project structure illustrated in Fig. 1 can be derived for the project model. As shown in Fig. 1, the project model does not make detailed restrictions regarding the project structure. There are only two requirements to be met by projects complying with the structure shown in Fig. 1 and thus with the general project model. First, the project needs to be defined under a given IT agenda. Second, the project must foresee development and operation of one or more pilot applications.

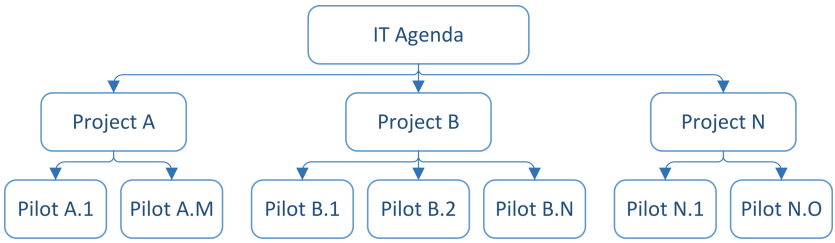

Fig. 1. The proposed evaluation scheme can be applied to all projects complying with the shown general structure (adapted from [17]).

By intention, the project structure of the general project model has been kept as abstract as possible. This assures a broad applicability of the general project model to concrete research projects. This, in turn, guarantees that the proposed evaluation scheme, which builds on the general project model defined in this section, can be applied to a broad set of research projects as well.

3.2 Project-Execution Phases

General project-execution phases constitute the second aspect of the general project model defined in this section. Similar to the project structure introduced in Sect. 3.1, also general project-execution phases can be derived from the three basic assumptions made above. Overall, the project model foresees three basic project-execution phases. In parallel, three evaluation phases are foreseen as well, covering evaluation-related tasks throughout the project.

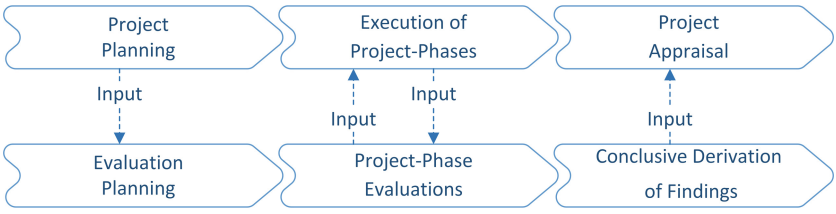

Fig. 2. The proposed evaluation scheme can be applied to all projects implementing the shown general project-execution phases (adapted from [17]).

Project-execution phases and related evaluation phases are shown in Fig. 2. According to this figure, the general project model considers the following three project-execution and evaluation phases:

- *Project Planning / Evaluation Planning*: During the Project Planning phase, the project's overall structure, contents, goals, and setup are defined. In parallel, the Evaluation Planning phase is executed, in which aspects related to evaluation activities within the project are defined. As shown in Fig. 2, relevant input from project planning needs to be considered during evaluation planning.
- *Execution of Project Phases / Project-phase Evaluations*: These two phases are executed after completion of the Project Planning and Evaluation Planning phases. According to the underlying assumptions of the general project model, the project is executed in consecutive project phases. Figure 2 shows that each project phase is accompanied by a corresponding evaluation phase. Figure 2 also shows that corresponding project-execution phases and evaluation phases influence each other. While conducted evaluations certainly depend on the respective project-execution phase and its goals and contents, project phases also should take into account available evaluation results (e.g. from previous phases). This supports a continuous improvement of the entire project.
- *Project Appraisal / Conclusive Derivation of Findings*: These phases are executed after completion of the last iterative project phase and corresponding project-phase evaluation. Hence, these phases are carried out at the end of the project to collect all lessons learned, draw the correct conclusions from these lessons, and to bring the project to a round figure. In the corresponding evaluation phase, findings are derived from conducted evaluations and serve as direct input to project appraisal. Figure 2 shows that the *Project Appraisal* phase needs to consider input from the *Conclusive Derivation of Findings* phase.

Similar to the project structure introduced in Sect. 3.1, also general project-execution phases have been defined on a rather abstract level. Together with the defined project structure, this yields a rather abstract general project model. As the general project model will serve as basis for the proposed evaluation scheme, its abstract nature assures that the proposed scheme is applicable to the majority of pilot-based projects targeting the public sector. Thus, the project model introduced in this section reasonably handles the trade-off between assuring general applicability and obtaining meaningful evaluation results.

4 Proposed Evaluation Scheme

Based on the defined general project model, we propose a generic evaluation scheme for the systematic evaluation of pilot-based public-sector research projects. The proposed evaluation scheme aims for two goals. On the one hand, the scheme aims to be abstract enough to be applicable to a broad range of pilots

and projects, in order to ensure comparability within projects (i.e. between the project's pilots) and also on a higher level between different projects. On the other hand, the evaluation scheme should still enable in-depth evaluations that take into account specifics of pilots and projects. The proposed scheme deals with this obvious trade-off by following a hierarchical approach.

In general, the evaluation scheme introduced in this section relies on the general project model defined above. If a research project complies with this general model, the proposed evaluation scheme defines in detail how to carry out evaluation-related activities throughout the project.

Essentially, the proposed evaluation scheme is composed of two building blocks. The first building blocks specifies an evaluation-criteria model. Evaluation criteria are crucial for any type of evaluation. Choosing adequate evaluation criteria turns out to be a challenging task, especially with regard to the given trade-off between general applicability and desired in-depth evaluations. The evaluation-criteria model, which is part of the proposed evaluation scheme, assists in overcoming this challenge by providing a framework for the definition of adequate evaluation criteria. Details of the proposed scheme's evaluation-criteria model are provided in Subsect. 4.1.

The second building block of the proposed evaluation scheme specifies the detailed evaluation process, which represents a step-by-step procedure to evaluate pilots using evaluation criteria that have been defined using the provided evaluation-criteria model. The proposed evaluation process builds up on the general project model introduced in Sect. 3 and further refines the various project and evaluation phases defined by this model. Details of the proposed scheme's detailed evaluation process are presented in Subsect. 4.2.

4.1 Evaluation-Criteria Model

The evaluation-criteria model constitutes the first major building block of the proposed evaluation scheme. It provides a framework that supports evaluators in defining and structuring evaluation criteria that are used during evaluation processes.

In principle, evaluation criteria used for pilot evaluation need to meet the same requirements as the overall evaluation scheme. Concretely, evaluation criteria should ideally be the same for all pilots in all evaluated projects. Only in this case, direct comparisons between different pilots and even between different projects are feasible. At the same time, evaluation criteria should be concrete enough to consider specifics of different pilots and projects. Obviously, these are contradictory requirements, which cannot be met by a simple one-dimensional list of evaluation criteria. Therefore, the proposed evaluation scheme follows a more complex approach and relies on a hierarchical evaluation-criteria model. This hierarchical model defines multiple layers, to which evaluation criteria can be assigned to. This is illustrated in Fig. 3.

The evaluation-criteria model shown in Fig. 3 considers the fact that evaluation criteria are typically closely related to project and pilot goals. Concretely, evaluation criteria are used to assess a pilot's or project's compliance

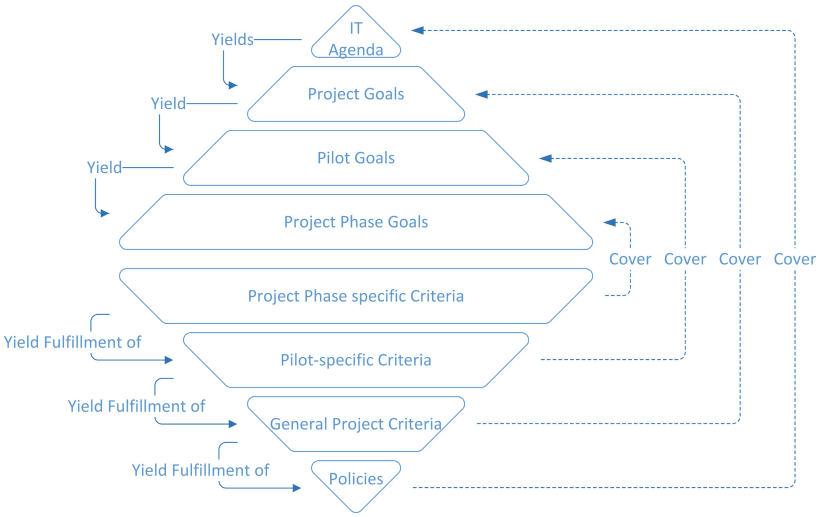

Fig. 3. The proposed evaluation-criteria model enables the definition of goals and related criteria on multiple layers of abstraction (adapted from [17]).

with defined goals. This close relation is also reflected by the evaluation-criteria model depicted in Fig. 3. The upper pyramid shows the different layers, on which relevant goals can be defined. Note that the pyramid's structure complies with the general project model defined in Sect. 3. The pyramid's topmost level represents the relevant IT agenda defining the very basic goals to consider. This agenda yields project-specific goals for concrete projects executed under this agenda. Within a concrete project, pilot-specific goals can be derived for each of the project's pilots from the overall project goals. Finally, pilot-specific goals can be further detailed by defining pilot-specific goals separately for each project phase. Overall, the proposed evaluation-criteria model defines relevant goals to be defined on four different layers of abstraction.

Once all goals have been defined according to the four layers, relevant evaluation criteria can be derived. The proposed model foresees evaluation criteria to be defined on four layers of abstraction as well, yielding the lower pyramid shown in Fig. 3. When deriving evaluation criteria for the four layers, two aspects need to be considered. First, defined evaluation criteria must cover the relevant goals defined before. This applies to all layers and is indicated in Fig. 3 by arrows between the upper and the lower pyramid. Second, evaluation criteria defined in neighboring layers must be related. In particular, given the fulfillment degree of criteria in one layer, it must be possible to derive the fulfillment degree of criteria in the subjacent layer.

Together the different layers for project-related goals and for evaluation criteria constitute the evaluation-criteria model shown in Fig. 3. Note that the proposed evaluation-criteria model deliberately does not define concrete evaluation criteria, in order to ensure a broad applicability of the proposed evaluation

model. Instead, the evaluation-criteria model merely defines a framework for the definition and classification of relevant goals and for the derivation of related evaluation criteria. The provided framework enforces a systematic derivation process for relevant goals and evaluation criteria on different layers of abstraction, supports the modeling of relations between goals and criteria defined on different layers, and assures a precise mapping between goals and evaluation criteria. This enables a systematic evaluation process, where the fulfillment of higher-level criteria can be derived automatically from the fulfillment of lower-level requirements.

4.2 Evaluation Process

Once relevant goals have been defined and evaluation criteria have been derived using the evaluation-criteria model introduced above, the actual evaluation process can be carried out. This evaluation process constitutes the second building block of the proposed evaluation scheme and is illustrated in Fig. 4. In principle, the shown process can be regarded as a detailing of the general project and evaluation phases introduced in Sect. 3. Thus, the proposed evaluation process implicitly complies with the defined general project model introduced in that section.

Figure 4, which illustrates the proposed evaluation process by means of a flow chart, is subdivided into six areas. First, the entire chart is split into two halves. The left half describes process steps to be carried out during project execution. Hence, this half of the flow chart corresponds to the upper part of Fig. 2. In contrast, the right half describes necessary steps to be carried out during project evaluation. Accordingly, this half of the chart corresponds to the lower part of Fig. 2. Second, the two halves of the flow chart are further split into three horizontal areas. This reflects the three general execution phases defined in Sect. 3. In summary, this yields six areas, each representing a project or evaluation phase, and each containing relevant process steps to be carried out in that phase.

Figure 4 shows that the proposed evaluation scheme comprises 18 processing steps to be carried out in total. For the first project-execution and evaluation phase, i.e. *Project Planning* and *Evaluation Planning*, the evaluation scheme foresees execution of the following steps:

- **Step (1) - Identification of Relevant IT-agenda Goals:** In this first step in project planning, general goals from relevant IT agendas, under which the project is executed, are identified. This step yields the most high-level goals to be considered during project execution. In the end, the overall success of the project is assessed against the goals identified in this step.
- **Step (2) - Definition of Project Goals:** From the general goals identified in Step (1), concrete project goals are derived in this step. Project goals detail higher-level goals by applying the respective project's specific context to them. Hence, the project's context and its defined contents are a relevant input during execution of this processing step.

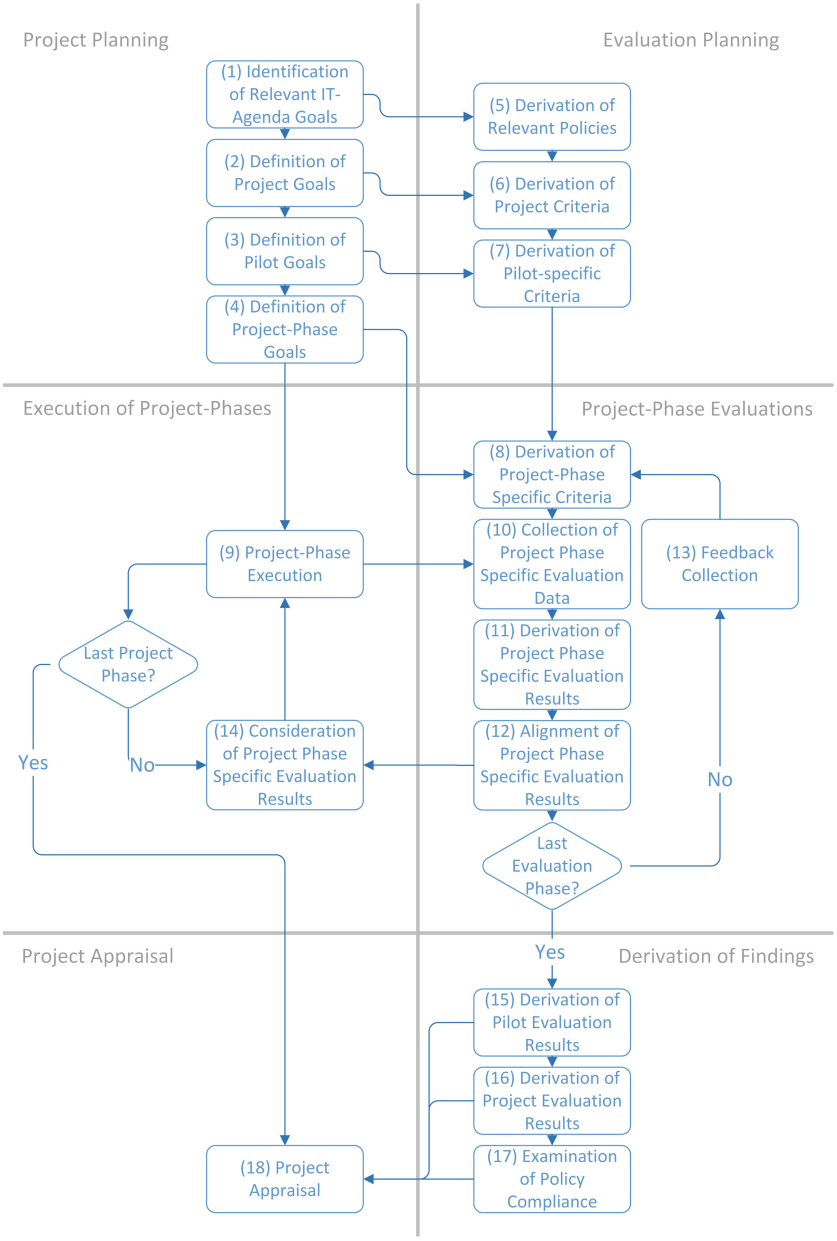

Fig. 4. The evaluation scheme specifies process steps to be carried out in the shown order (adapted from [17]).

– **Step (3) - Definition of Pilot Goals:** Once the overall project goals have been fixed in Step (2), pilot-specific goals need to be derived. Pilot-specific goals of course need to comply with the higher-level project goals, but

additionally take into account the specifics of the project's different pilot applications. Accordingly, the definition of pilots and their foreseen role in the project are relevant inputs to this processing step.

- **Step (4) - Definition of Project-phase Goals:** Pilot-specific goals obtained in Step (3) can again vary between different project phases. For instance, concrete goals of a pilot can vary between project phases, in which the pilot is developed, and phases, in which the pilot is operated. To consider possible phase-specific variations, pilot-specific goals are further detailed to project phase specific goals in this step.
- **Step (5) - Derivation of Relevant Policies:** This is the very first step to be taken in evaluation planning. In this step, relevant policies to be considered are derived from general goals extracted from relevant IT agendas. Hence, results of Step (1) are a relevant input to this step. In the end, the overall evaluation process will tell whether the project complies with the policies identified in this step.
- **Step (6) - Derivation of Project Criteria:** By taking into account the policies derived in Step (5) and the general project goals obtained in Step (2), evaluation criteria are defined on project level in this step. These criteria must be defined such that their degree of fulfillment can tell whether the project meets its goals defined in Step (2).
- **Step (7) - Derivation of Pilot-specific Criteria:** From the project criteria obtained from the previous step, pilot-specific criteria are derived in this step. Pilot-specific goals obtained in Step (3) are a relevant input to this step. Derived pilot-specific criteria must be defined such that their degree of fulfillment can tell whether the respective pilot meets its goals defined in Step (3).

After completing the process steps described so far, all relevant goals and (almost) all evaluation criteria have been defined and set in relation according to the evaluation-criteria model introduced in Sect. 4.1. What is still left to be done is the derivation of project phase specific evaluation criteria (Step (8)). This task is intentionally shifted to the subsequent project-execution/evaluation phase (i.e. the next horizontal area shown in Fig. 4), as the proposed evaluation model foresees a dynamic and iterative adaption of these criteria during the entire project life-cycle. Overall, the next project-execution and evaluation phase (i.e. *Execution of Project Phases* and *Project Phase Evaluations*) comprise the following processing steps:

- **Step (8) - Derivation of Project-phase Specific Criteria:** In this step, project phase specific evaluation criteria are derived for each pilot, taking into account the respective pilot's pilot-specific criteria from Step (7) on the one hand, and project phase specific goals derived in Step (4) on the other hand. For all but the first project phase, feedback collected during the previous project phase is another relevant input to this step and must be considered for the definition of criteria for the current project phase. This way, project phase specific criteria can be dynamically adapted during the entire project life-cycle to consider changing circumstances and requirements.

- **Step (9) - Project-phase Execution:** This step covers the execution of a project phase according to the project setup. Note that this is a rather complex and time-consuming step in most cases, as it covers all project-related activities assigned to a specific project phase. However, details of this processing step are usually irrelevant for the proposed evaluation scheme. Hence, these details are not modeled in the flow chart shown in Fig. 4.
- **Step (10) - Collection of Project Phase Specific Evaluation Data:** This step is executed in parallel to Step (9), i.e. in parallel to the execution of a certain project phase. During this step, evaluation data is collected. This can be achieved e.g. through interviews, questionnaires, or through the automated measuring or logging of data. The proposed evaluation scheme leaves the choice of the most suitable method to the respective project evaluators. This ensures a broad applicability of the proposed evaluation scheme, as the best tool choice can depend on the nature of the project to be evaluated. In any case, Step (9) provides relevant input to this processing step.
- **Step (11) - Derivation of Project Phase Specific Evaluation Results:** In this step, evaluation data collected in Step (10) are analyzed in order to derive evaluation results for the current project phase. The concrete analysis process depends on the type of evaluation data collected. Hence, the proposed evaluation scheme again does not apply any restrictions here. In any case, project-phase specific criteria derived in Step (8) are a prerequisite for this step.
- **Step (12) - Alignment of Project Phase Specific Evaluation Results:** Evaluation results of the current project phase obtained in Step (11) are subsequently aligned with other relevant stakeholders involved in the project. Depending on the respective project setup, this can be e.g. pilot developers or pilot operators. This step gives involved stakeholders the chance to comment on results obtained, in order to prevent misunderstandings early and to ensure a consensual overall evaluation process.
- **Step (13) - Feedback Collection:** In case there is at least one more project phase (and hence also one more evaluation phase) to come, feedback on the applied evaluation process is collected from all involved stakeholders. Again, the method applied to collect feedback is left open to the respective evaluators. Also in this context, the best choice can depend on the actual project setup. Independent of the chosen method, collected feedback serves as input for the definition of project phase specific criteria for the next project phase (Step 8). If no further project phase is foreseen, i.e. the project has just completed its last phase, this step is omitted.
- **Step (14) - Consideration of Project Phase Specific Evaluation Results:** After completion of a project phase, and if there is another project phase to come, this step is carried out by the project executors. In this step, evaluation results of the just completed project phase are analyzed. If possible, lessons learned are derived and considered as input for the execution of the next project phase. This way, the project is continuously improved based on intermediate evaluation results. In case the last project phase has already been completed and no further phase is to come, this step is omitted.

After completion of the final project phase, the last two overall phases, i.e. *Project Appraisal* and *Derivation of Findings* are carried out. Relevant processing steps of these phases are described in the following:

- **Step (15) - Derivation of Pilot Evaluation Results:** This step is executed after completing the last iterative project and evaluation phases. In this step, project phase specific evaluation results collected iteratively in Step (10) and analyzed in Step (11) are combined for each pilot to derive overall pilot-specific evaluation results. If evaluation criteria have been defined according to the proposed evaluation-criteria model, this step can be carried out efficiently, as pilot-specific evaluation results can be derived directly from project phase specific results. After completion of this step, project evaluators know if and to what extent pilot-specific criteria defined at the beginning of the project in Step (7) are met.
- **Step (16) - Derivation of Project Evaluation Results:** Once all pilot-specific evaluation results have been derived in Step (15), this step combines them to overall project evaluation results. Again, this is an efficient process, if evaluation criteria have been defined such that the degree of fulfillment can be derived from the lower layer of the evaluation-criteria model. After completion of this step, project evaluators know if and to what extent project criteria defined in Step (6) are met.
- **Step (17) - Examination of Policy Compliance:** This step finally checks derived project evaluation results against relevant policies identified at the beginning of the project (Step 5). This way, stakeholders can assess whether the project and its results comply with IT agendas, from which these policies have been derived. This process step concludes evaluation-related activities within the project. All evaluation results (pilot evaluation results, project evaluation results, and policy compliance) serve as input for the final project-appraisal phase (Step 18).
- **Step (18) - Project Appraisal:** In this final step, all evaluation results stemming from different layers of abstraction are combined to derive the most valuable lessons learned. Due to the multi-layered approach followed, detailed analyses of evaluation results can be conducted, including specific results as well as comparisons between pilots and even between different projects.

4.3 Configurability

Although the proposed evaluation scheme specifies in detail necessary steps to be carried out, it deliberately remains flexible and abstract in certain aspects and gives evaluators, who apply the scheme in practice, room for configuration. This makes the proposed evaluation scheme more flexible and applicable to a broader range of projects. In particular, it ensures that the evaluation scheme can be applied to all pilot-based research projects complying with the general project model introduced in Sect. 3. Concretely, the proposed evaluation scheme can be configured and adapted to the respective project, to which it is to be applied, by means of the following aspects:

- **Number of Iterative Project Phases:** The evaluation scheme supports an arbitrary number of iterative project phases and corresponding evaluation phases. Not that this does not require the project to define multiple project phases. The evaluation scheme works with one single project phase (and corresponding evaluation phase) as well. However, it is not limited to that but supports 1 to N phases.
- **Evaluation Criteria:** The proposed evaluation scheme comprises an evaluation-criteria model. While this model provides a hierarchical framework for the definition and classification of evaluation criteria, it does not define concrete evaluation criteria. Instead, concrete evaluation criteria can be defined by the respective project evaluators, taking into account the specifics of the project and pilots to be evaluated.
- **Relation between Evaluation Criteria:** The proposed evaluation-criteria model enables the definition of evaluation criteria on multiple layers of abstraction. Criteria on neighboring layers should be set in relation to each other. This way, the fulfillment degree of higher-level criteria can be derived systematically from the fulfillment degree of lower-level criteria. The model does not impose any restrictions regarding the definition of relations between criteria. For instance, evaluators can follow a rather simple approach, which assumes each criterion to be equally important. Alternatively, more complex relations can be defined between different layers of the evaluation-criteria model. While this enables a more fine-grained modeling of dependencies between criteria on different layers of abstraction, it potentially complicates the derivation of evaluation results. However, evaluators are free to choose the approach that fits best the requirements of the respective project.
- **Method for Evaluation-Data Collection:** In each iterative project phase, evaluation data needs to be collected by evaluators. In principle, different approaches can be followed to carry out this task. For instance, data can be collected with the help of questionnaires, by means of personal interviews, by doing project-internal audits, or by collecting and analyzing generated data such as log files. The proposed evaluation scheme is flexible enough to support all these (and also other) approaches. Hence, evaluators can choose the method that fits best the characteristics of the project to be evaluated.
- **Method for Deriving Evaluation Results from Evaluation Data:** In each iterative project phase, evaluation results need to be derived from collected evaluation data. The proposed evaluation scheme does not specify this step and the methods to be applied in this step in detail. Hence, it is up to the executing evaluators to agree on an appropriate method that enables comprehensible derivation of evaluation results from collected evaluation data. Of course, the method applied needs to be aligned with the type of collected evaluation data. For instance, if this data is available in structured form, automated methods can be applied to derive evaluation results from it. In case evaluation data has been collected e.g. by means of telephone interviews, manual methods may be necessary.
- **Method for Alignment of Project Phase Specific Evaluation Results:** The proposed evaluation scheme gives involved stakeholders such as

pilot developers or pilot operators the opportunity to comment on intermediate evaluation results in each project and corresponding evaluation phase. However, the scheme does not define a concrete method to be applied for carrying out this task. Also for this aspect of the evaluation process, evaluators are free to choose their preferred approach to align intermediate evaluation results with other involved stakeholders.

– **Provision of Feedback:** Finally, the proposed evaluation scheme also defines a feedback channel from involved stakeholders to the project evaluators. Again, the scheme does not define a concrete implementation of this feedback channel but leaves this decision to the executing evaluators. This way, evaluators can choose the approach fitting best the characteristics of the respective project.

Of course, leaving certain aspects unspecified imposes an additional task on the evaluators of a project. They need to parametrize the evaluation scheme, in order to adapt it to the specifics of the respective project. However, we believe the gain in flexibility and applicability definitely compensates for the additional effort. In the following section, we show one possible parametrization by discussing the application of the proposed evaluation scheme to the EU-funded research project FutureTrust and its pilots.

5 Evaluation

Future Trust Services for Trustworthy Global Transactions (FutureTrust) is an international research project funded by the EU under the programme *H2020-EU.3.7. - Secure societies - Protecting freedom and security of Europe and its citizens*. The project consortium consists of 16 partners from 10 countries, including EU member states as well as third-party countries. The overall aim of FutureTrust is to support the practical implementation of the EU eIDAS Regulation [8] in Europe and beyond. Software components developed by FutureTrust are applied to real-world use cases by means of several pilots and demonstrators.

FutureTrust fully complies with the general project model described in Sect. 3. The project has a focus on the public-sector domain and is motivated by an EU agenda, as described in the project's funding programme [6]. Furthermore, FutureTrust develops and operates a series of demonstrators and pilots. Thus, the evaluation scheme proposed in this article is well suited for carrying out pilot evaluations in FutureTrust. First experiences gained during application of the proposed evaluation scheme in FutureTrust are discussed in this section.

5.1 Configuration

The proposed evaluation scheme introduced in Sect. 4 intentionally remains generic in several aspects and hence needs to be configured and parametrized before being applied to a concrete research project. In the case of FutureTrust, the following parameters have been chosen for configurable aspects listed in Sect. 4.3:

- **Number of Iterative Project Phases:** In principle, the evaluation scheme supports an arbitrary number of iterative project phases and corresponding evaluation phases. For FutureTrust, three phases have been defined. Evaluations are carried out before piloting (ex-ante evaluation), during piloting (mid-term evaluation), and after piloting (ex-post evaluation). This complies with FutureTrust's overall project setup as defined in FutureTrust's project description.
- **Evaluation Criteria:** While the proposed evaluation scheme provides a hierarchical evaluation-criteria model for the definition and classification of evaluation criteria, it does not define concrete evaluation criteria. For FutureTrust, relevant evaluation criteria have been extracted from relevant agendas and project-definition documents. Details on chosen evaluation criteria are provided in Subsect. 5.2.
- **Relation between Evaluation Criteria:** According to the proposed evaluation-criteria model, evaluation criteria assigned to neighboring layers of abstraction should be set in relation to each other. This way, the fulfillment degree of higher-level criteria can be derived systematically from the fulfillment degree of lower-level criteria. The model does not impose any restrictions regarding the definition of relations between criteria. For the sake of simplicity, we have followed a rather simple approach in FutureTrust. This approach assumes each criterion to be equally important and to have the same impact on related criteria in neighboring layers of abstraction.
- **Method for Evaluation-data Collection:** The proposed scheme requires evaluators to collect evaluation data in each iterative project phase. As FutureTrust piloting partners are distributed all over Europe, questionnaires have been prepared and sent out to piloting partners to collect necessary evaluation data. Where necessary, telephone calls have been organized to clarify open issues in a bilateral way.
- **Method for Deriving Evaluation Results from Evaluation Data:** In each iterative project phase, evaluation results need to be derived from collected evaluation data. In FutureTrust, evaluation data has been collected by means of questionnaires. As these questionnaires contain answers in prose, i.e. in unstructured form, an automated processing and analysis of evaluation data has not been feasible. Thus, the analysis of questionnaires and the derivation of intermediate evaluation results has relied on manual processes.
- **Method for Alignment of Project Phase Specific Evaluation Results:** The proposed evaluation scheme gives all involved stakeholders the opportunity to continuously comment on intermediate evaluation results. Due to the local dispersion of stakeholders within the project, FutureTrust follows again a document-based approach for this task. Evaluation results are compiled into intermediary evaluation reports. These reports are sent out to all involved stakeholders in order to give them the opportunity to provide feedback in written form.
- **Provision of Feedback:** The proposed evaluation scheme defines a feedback channel from involved stakeholders to the evaluators. FutureTrust organizes regular meetings (online and face-to-face) that bring together involved

stakeholders. These meetings can be used to bilaterally provide feedback on the overall evaluation process as defined by the proposed evaluation scheme.

5.2 Results

After applying the configuration as described in Sect. 5.1, the proposed evaluation scheme has been applied to the FutureTrust project and its pilots. In particular, the scheme has been applied twice, once for the evaluation of FutureTrust pilots and demonstrators, and once for analyzing their impact. The applied evaluation process has been exactly the same for pilot evaluation and impact analysis. However, different (hierarchical) evaluation criteria have been defined to reflect the different scopes of pilot evaluation and impact analysis.

Based on EU agendas relevant for FutureTrust, the following general project criteria have been defined for pilot evaluation:

- *Security and Data Protection*
- *Functionality*
- *Usability*
- *Interoperability*
- *Reusability and Sustainability*
- *Legal Compliance*
- *Compliance with Project Goals*

Accordingly, the following general project criteria have been defined for impact analysis:

- *Demonstration of Positive Business Case*
- *Empowerment and Protection of Users*
- *Increase of Use of Trust Services*
- *Reduction of Administrative Overhead*
- *Adherence to Sufficient Technology Readiness Level (TRL)*

From these general project criteria, pilot-specific evaluation criteria have been derived for each FutureTrust pilot and demonstrator. For each pilot and demonstrator, 24 pilot-specific criteria have been defined for pilot evaluation. In addition, 9 pilot-specific criteria have been defined for impact analysis for each pilot and demonstrator. Finally, project phase specific evaluation criteria have been derived. In total, 33 criteria for pilot evaluation and 9 criteria for impact analysis. To assure meaningful evaluation results, pilot-specific evaluation criteria have been defined based on the SMART approach described by Doran [3]. Accordingly, all defined criteria are **S**pecific, **M**easurable, **A**chievable, **R**elevant, and **T**ime-bound.

Based on the derived project phase specific evaluation criteria, questionnaires have been prepared and sent out to pilot developers and operators. By analyzing returned filled questionnaires, the fulfillment degree of project phase specific evaluation criteria could be determined. Where necessary, ambiguities have been clarified in bilateral calls.

From the obtained project phase specific evaluation results, the fulfillment of higher-layer criteria could be derived automatically. As an illustrative example, Figs. 5 and 6 show first results of the conducted pilot evaluation and impact analysis. Obtained results demonstrate the applied evaluation scheme's capability to yield evaluation results that are comparable among different FutureTrust pilots.

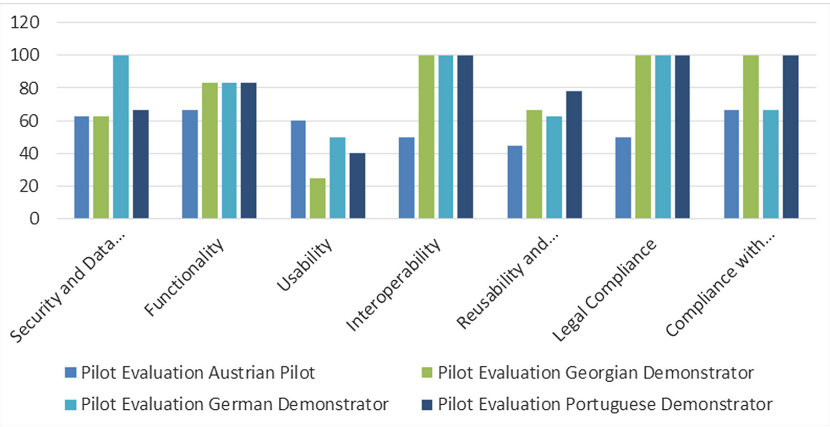

Fig. 5. Results of FutureTrust pilot evaluation (adapted from [17]).

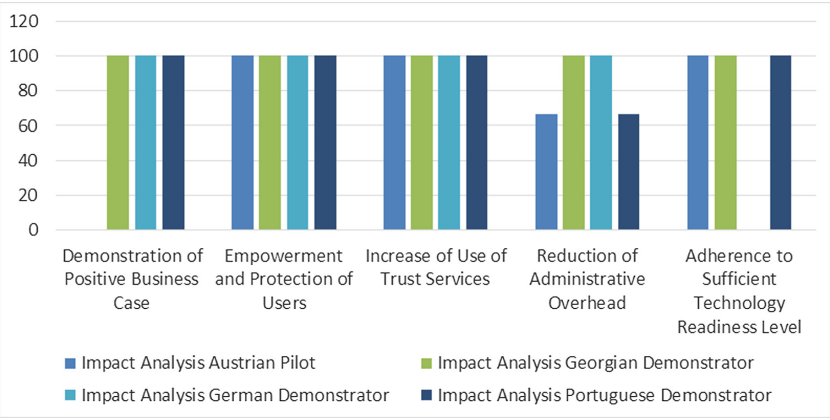

Fig. 6. Results of FutureTrust impact analysis (adapted from [17]).

Overall, the successful completion of the described evaluation steps in the context of FutureTrust demonstrates the practical applicability of the proposed evaluation scheme. The FutureTrust project was still ongoing when obtaining the evaluation results described above. Accordingly, the entire project evaluation

was not finished at that time. However, most processing steps of the proposed evaluation scheme (including the most challenging ones like the definition of evaluation criteria) could already be applied and completed successfully. The proposed scheme can hence be regarded suitable for application in practice.

6 Lessons Learned

Although the FutureTrust project was not completed at the time of writing this article, several useful findings and lessons learned could be obtained during a first application of the proposed evaluation scheme to FutureTrust pilots and demonstrators in the first project phases.

The probably most important lesson learned is the finding that the proposed scheme can by applied successfully to a concrete research project. All concepts and processing steps defined by the proposed scheme could be applied and carried out as foreseen. Furthermore, the proposed scheme is able to meets its basic goal, i.e. yield comparable evaluation results while still taking into account specific characteristics and properties of different pilots.

Another positive lesson learned is the finding that the proposed scheme is rather flexible with regard to changes in the overall project structure. As pilots are treated independently to a large extent, varying schedules of different pilots caused by unexpected delays, or even the inclusion of new pilots during the project lifetime does not cause severe problems.

A less positive lesson learned concerns the scheme's evaluation-criteria model. It turned out that some assumptions made when designing this model were too optimistic. Concretely, the proposed evaluation scheme assumes that pilot-specific criteria are defined once at the beginning of the project. Although project phase specific criteria are derived from these pilot-specific criteria in each project phase, the pilot-specific criteria are assumed to be stable. FutureTrust has shown that this is not always the case in practice. Indeed, a minor extension of pilot-specific criteria was necessary after the first project phase, in order to achieve necessary alignments with other evaluation-related tasks in the project. While this was no show stopper in practice, the evaluation scheme should probably consider such necessities.

Overall, first lessons learned from applying the proposed evaluation scheme in FutureTrust are predominantly positive. While minor room for improvement has been identified, the scheme has passed its practical test so far. FutureTrust will continue to rely on the proposed evaluation scheme for the remaining project lifetime to obtain valuable evaluation results.

7 Conclusions and Future Work

In this article, we have proposed and introduced an evaluation scheme for pilot-based research projects. Based on the awareness that project evaluations of

past research projects have not always been as good as they could be, the proposed scheme aims to support the systematic derivation of both comparable and detailed evaluation results.

To meet this apparently contradictory goal, the proposed scheme follows a hierarchical approach. Concretely, it defines evaluation criteria on different layers of abstraction and provides a detailed process description for carrying out evaluations against these criteria. Its successful application within an international research project shows that the proposed scheme indeed is able to meet its goal. The scheme's hierarchical nature enables in-depth evaluations of specific pilots, while still guaranteeing comparability of obtained evaluation results on a higher level of abstraction. Ultimately, the proposed scheme yields more valuable evaluation results, from which all stakeholders involved in a research project can benefit in the end. Especially funding bodies such as the EU can take advantage of more homogenous evaluation results, which supports them in steering research activities into the right directions, and in complying with relevant IT agendas.

Lessons learned from applying the proposed evaluation scheme to the research project FutureTrust still show some room for improvement. We continuously consider new findings to further improve the scheme. Applied improvements are continuously tested and evaluated by applying the scheme in the remaining project phases of FutureTrust. For the future, we also plan to apply the scheme to other projects, to further evaluate its project-independent applicability and to test the comparability of evaluation results obtained from different projects. At the same time, we aim to extend the scope of the proposed evaluation scheme. While its current focus lies on pilot-based projects from the public sector, we plan to make it applicable to other project types as well. This way, an even broader set of research projects could benefit from the proposed scheme.

Acknowledgments. This work has been supported by the FutureTrust project (N.700542) funded under the programme H2020-EU.3.7. - Secure societies - Protecting freedom and security of Europe and its citizens (2013/743/EU of 2013-12-03).

References

1. Asosheh, A., Nalchigar, S., Jamporazmey, M.: Information technology project evaluation: an integrated data envelopment analysis and balanced scorecard approach. Exp. Syst. Appl. **37**(8), 5931–5938 (2010). https://doi.org/10.1016/j.eswa.2010.02.012. http://www.sciencedirect.com/science/article/pii/S0957417410000515
2. De Cock, D., Wouters, K., Preneel, B.: Introduction to the Belgian EID card. In: Katsikas, S.K., Gritzalis, S., López, J. (eds.) EuroPKI 2004. LNCS, vol. 3093, pp. 1–13. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-25980-0_1
3. Doran, G.T.: There's a SMART way to write managements's goals and objectives. Manag. Rev. **70**(11), 2 (1981). EBSCOhost. http://web.a.ebscohost.com/ehost/pdfviewer/pdfviewer?sid=c62fb711-f4fb-4a4e-9f6e-62de080316e0%40sessionmgr4006&vid=1&hid=4101

4. Eilat, H., Golany, B., Shtub, A.: R&d project evaluation: an integrated dea and balanced scorecard approach. Omega **36**(5), 895–912 (2008). https://doi.org/10.1016/j.omega.2006.05.002. http://www.sciencedirect.com/science/article/pii/S0305048306000442
5. European Commission: Digital Agenda for Europe (2018). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=LEGISSUM:si0016&from=DE. Accessed 05 July 2018
6. European Commission: H2020-EU.3.7. - Secure societies - Protecting freedom and security of Europe and its citizens (2018). https://cordis.europa.eu/programme/rcn/664463_en.html. Accessed 05 July 2018
7. European Commission: Shaping the Digital Single Market (2018). https://ec.europa.eu/digital-single-market/en/policies/shaping-digital-single-market. Accessed 05 July 2018
8. European Union: Regulation (EU) No 910/2014 of the European Parliament and of the Council of 23 July 2014 on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC (2018). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32014R0910. Accessed 05 July 2018
9. Khan, Z., Ludlow, D., Caceres, S.: Evaluating a collaborative it based research and development project. Eval. Prog. Planning **40**, 27–41 (2013). https://doi.org/10.1016/j.evalprogplan.2013.04.004. http://www.sciencedirect.com/science/article/pii/S0149718913000347
10. Knall, T., Tauber, A., Zefferer, T., Zwattendorfer, B., Axfjord, A., Bjarnason, H.: Secure and privacy-preserving cross-border authentication: the STORK pilot 'SaferChat'. In: Andersen, K.N., Francesconi, E., Grönlund, Å., van Engers, T.M. (eds.) EGOVIS 2011. LNCS, vol. 6866, pp. 94–106. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22961-9_8
11. Leitold, H., Hollosi, A., Posch, R.: Security Architecture of the Austrian Citizen Card Concept. In: 18th Annual Computer Security Applications Conference, 2002. Proceedings, pp. 391–400 (2002). https://doi.org/10.1109/CSAC.2002.1176311
12. Leitold, H., Posch, R.: STORK - technical approach and privacy. In: Bus, J., Crompton, M., Hildebrandt, M., Metakides, G. (eds.) Digital Enlightenment Yearbook 2012, pp. 289–306 (2012)
13. Martens, T.: Electronic identity management in Estonia between market and state governance. Identity Inf. Soc. **3**(1), 213–233 (2010). https://doi.org/10.1007/s12394-010-0044-0
14. Rath, C., Roth, S., Bratko, H., Zefferer, T.: Encryption-based second authentication factor solutions for qualified server-side signature creation. In: Kő, A., Francesconi, E. (eds.) EGOVIS 2015. LNCS, vol. 9265, pp. 71–85. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22389-6_6
15. Suzic, B., Reiter, A.: Towards secure collaboration in federated cloud environments. In: 2016 11th International Conference on Availability, Reliability and Security (ARES), pp. 750–759, August 2016. https://doi.org/10.1109/ARES.2016.46
16. Tauber, A., Zwattendorfer, B., Zefferer, T.: Stork: pilot 4 towards cross-border electronic delivery. In: Verlag, T. (ed.) Electronic Government and Electronic Participation - Joint Proceedings of Ongoing Research and Projects of IFIP EGOV and ePart 2011, pp. 295–301 (2011)
17. Tou, J.T.: Information systems. In: von Brauer, W. (ed.) GI 1973. LNCS, vol. 1, pp. 489–507. Springer, Heidelberg (1973). https://doi.org/10.1007/3-540-06473-7_52