PHENICX

D2.8. Evaluation methodology report, version 1

Grant Agreement nr	601166
Project title	Performances as Highly Enriched aNd Interactive Concert eXperiences
Project acronym	PHENICX
Start date of project (dur.)	Feb 1st, 2013 (3 years)
Document reference	PHENICX-D-WP2-TUD-20130729-D2.8-Evaluation_methodology_v1-1.2
Report availability	PU – Public
Document due Date	Jul 31th, 2013
Actual date of delivery	Jul 29, 2013
Leader	TUD
Reply to	Cynthia Liem (TUD) (c.c.s.liem@tudelft.nl)
Additional main contributors (author's name / partner acr.)	Mark Melenhorst (TUD) Martha Larson (TUD) Marcel van Tilburg (RCO) Ron van der Sterren (VD)
Document status	Draft version (reviewer: JKU)

Project funded by ICT-7th Framework Program from the European Commission

PHENICX

Table of Contents

1 BACKGROUND

This deliverable, Deliverable 2.8 "Evaluation methodology report, version 1", has the goal of providing methodological guidelines for the development and evaluation that will take place in the PHENICX project. At month 31 of the project, this deliverable will be followed by Deliverable 2.16 "Evaluation methodology report, final version", which will go beyond the initial guidelines laid out in this document to give the details of the evaluation methodology.

2 INTRODUCTION

2.1 Main objectives and goals

Evaluation is the process of determining worth or usefulness with respect to a set of standards. In PHENICX, evaluation is applied to reach two basic objectives:

1. Gathering information that will allow us to **understand how PHENICX technologies can be improved**, and by how much they must be improved in order to make a difference to users (also known as *formative evaluation*).
2. Verifying that PHENICX has **achieved the targets for success** that it has set for itself (also known as *summative evaluation*).

This deliverable provides the first description of the methodology that will be used to carry out evaluation in the PHENICX project in the form of a set of guidelines. These guidelines will provide the basis for designing specific evaluations that will be carried out later in the project.

As defined in the DoW, technology developed in PHENICX takes the form of two different types of implementations:

1. The development of *individual prototypes* (WP3-5).
2. The development of *Integrated Prototype Systems*, in which multiple prototypes are combined (WP6).

For both prototypes and Integrated Prototype Systems, distinct evaluations will take place and distinct methodologies will be used. This deliverable treats each of these in turn.

Evaluation of individual prototypes in PHENICX takes place with respect to the success criteria that are set out in the project. In Deliverable 2.3 "Technical success criteria", an initial set of success criteria have been established as a first step towards the evaluation. Success criteria for the PHENICX prototypes fall into two classes: *algorithmic success criteria* (indicators for the success of algorithms independent of individual implementations) and *system and user success criteria* (indicators for the success of PHENICX technology that has been implemented into a prototype). Although user studies will be mainly carried out on Integrated Prototype Systems, rather than on prototypes, information gathered by users interacting with Integrated Prototype Systems can be projected down to individual system components and in this way used to gauge the usefulness of particular algorithms to users.

The integrated prototype systems serve two purposes:

1. Demonstrating PHENICX technologies in real-world applications for practical non-academic stakeholders.
2. Testing them with a realistic user audience.

This will be done as part of both the formative and the summative evaluation of the integrated prototype systems, as part of WP 7, Task 7.1 (Demonstrator development and testing) and Task 7.2 (Implementation and testing of applications in real use cases).

The development of Integrated Prototype Systems follows a *user-centered design approach*, in which each iteration in the software development process is followed by a round of user feedback. This approach enables us to learn about end-user needs, preferences, and user acceptance, relevant to formative evaluation. In the final evaluation, we focus on summative evaluation and will evaluate whether the success criteria are met overall.

The development of individual prototypes precedes the development of integrated prototype systems. End-users will also be involved during the development of prototypes. The following tasks will involve feedback from end-users:

- In *Task 6.1: Visualisation of music pieces and their performances* meaningful visualization tools will be developed, that will be subjected to evaluation with end-users, both in a qualitative and a quantitative way.
- In *Task 6.2: Personalised multimodal information system* the user will be provided with personalized pieces of music-related information. A small-scale qualitative survey will be conducted among potential users.
- In *Task 6.3: Acoustic rendering of augmented music performances* both perceptual experiments with end users and quantitative evaluations will be conducted.
- In *Task 6.4: Interactive systems for performer impersonation* an end-user evaluation will be conducted, focusing on the interaction between the end-user, his movements, and the effect of his or her movements.

2.2 Executive summary

This deliverable makes three main contributions:

- It sets out the **initial guidelines for evaluation methodologies** that will be used to evaluate the technical performance evaluation of the PHENICX algorithms and prototypes.
- It provides **methodological guidance for both the formative and summative evaluation of the PHENICX Integrated Prototype Systems**, including sets of indicators for each type of evaluation.
- It describes **end-user involvement strategies** to recruit users to carry out the evaluation of the PHENICX Integrated Prototype Systems.

2.3 Methodology

This deliverable is focused on providing methodological guidelines for project success evaluation at different levels. These guidelines largely follow out of earlier experiences of the contributing authors in benchmarking activities and international cooperation projects.

2.4 Terminology

PHENICX Prototypes
Implementations of individual technologies, usually representing the output of individual PHENICX tasks.

PHENICX Integrated Prototype Systems
Systems that integrated multiple PHENICX technologies that have been created by different tasks. The purpose of PHENICX Integrated Prototype Systems is to carry out demonstrations and texts with real-life users in real-world scenarios and contexts.

Formative evaluation

Activities that have the purpose of collecting feedback on and making improvements to a prototype or integrated prototype system. Several methods and techniques are available to accomplish this task. The object of evaluation has an increasing level of maturity, starting from 'paper' use cases and, in the end, working demonstrators.

Summative evaluation
Activities aimed at assessing whether the success criteria have been met during the final evaluation of the PHENICX integrated prototype systems.

Success criteria
Statements that make it possible to understand the point at which PHENICX has reached its goals and make it possible to measure and track improvement in *technical performance*.

User-based performance indicators
Statements that make it possible to understand the point at which PHENICX has reached its goals from the perspective of end-users. They are different from technical success criteria in the sense that no target levels can be specified as a point of reference is lacking.

2.5 Convention

We use the following writing conventions:

- **bold** for emphasis
- *italics* for newly introduced terminology
- underlined for cross-references and references to other documents

3 EVALUATION METHODOLOGY FOR ALGORITHMS & PROTOTYPES

Algorithms and prototypes are evaluated quantitatively using specific data sets. A key characteristic of a good evaluation is its **reproducibility**. In other words, a new research group should be able to take (or re-implement) an already-developed algorithm and obtain the same results when they evaluate it. Reproducibility guarantees that it is possible to establish a "baseline" performance level, in other words a level of performance which a new algorithm or prototype must improve upon in order to claim that it is extending the state of the art.

While a lot of the data and code used within PHENICX will not be publicly shareable, researchers are requested to still **make their work as reproducible and transparent as possible**. This can e.g. be established by clearly outlining steps that were taken towards a certain algorithm with their underlying reasoning, and by giving a clear overview of relevant tuning parameters that were chosen. Furthermore, if a work builds forth on an earlier work, it also is good practice to actively reuse and reproduce the main findings of the earlier work.

A key aspect of evaluation methodology is **making the right choice of metric**. The metric is the scoring function by which the performance is measured. For example, a common metric for measuring the quality of a result list is *precision*, i.e., the proportion of relevant results among the total number of results returned by the system. The metric provides the "perspective" on the performance that is "seen" by the evaluation. **It is critical that this perspective is correct.** For example, if a technology is to be used in a system in which it is known that users will look only at the topmost relevant result, measuring precision does not provide a useful perspective on the performance of that technology. Instead, the metric chosen should then be the so-called "precision at one", which reflects the relevance of only the topmost result.

For the reasons outlined above, **metrics used to evaluate PHENICX outcomes should be accompanied by a documented reasoning on why these particular metrics were chosen** (either in corresponding deliverables or publications).

We distinguish two basic types of evaluation methodology which will be used to evaluate algorithms and prototypes in the PHENICX project*: ground-truth-based evaluation*, under which algorithms will be run on a data set and compared with earlier established reference labels or annotations, and *output-based evaluation*, under which an algorithm is applied to a data set and human judges focus their consideration on the output of the algorithm.

Both evaluation methodologies require human involvement and have advantages and disadvantages. An advantage of ground-truth-based evaluation is that as soon as the ground truth is established, it can easily be reused when new algorithms or approaches are developed; a disadvantage is that it frequently requires human annotations of an entire data set, which is a time and human-resource consuming process. An advantage of output-based evaluation is that it can give insight into the results of an algorithm even when no indisputable ground truth can be established. However, the disadvantage of output-based evaluation is that each new algorithm will require a new round of annotation, restricting the ease with which new algorithms can be evaluated.

In the rest of this section, we give further details about both mentioned methodology types.

3.1 Ground-truth-based evaluation

For ground-truth-based evaluation to be most effective, it must **start from the basis of a use scenario** that details **how a particular algorithm or prototype will be used in the real world**. The PHENICX use cases, detailed in Deliverable 2.2, provide information about such scenarios. On the basis of a scenario, a proper *data set* can be identified that is representative of the data that would be used by users in the scenario.

Data set identification often involves making a trade-off between having a very large data set and having a data set that can be annotated by human judges given the time and resources available. In any case, **it is critical to ensure that the data set is sampled in a way that maintains the key characteristics of the data that is used in the use scenario**. For example, careful attention must be paid to maintaining *diversity*. Additional information about data set creation requirements in the PHENICX project are given in Deliverable 2.5, "Corpus generation guidelines, version 1".

Data sets should be divided into **training, development and test partitions.** *Development partitions* are used to tune parameters while *test partitions* are used to perform the actually evaluation. **Any evaluation that allows the test data to bleed into the tuning or training data will not yield a valid evaluation result.**

Once the data set has been established, it is annotated with reference labels, either by human judges or by an automated procedure. These reference labels are referred to as the *ground-truth* because they reflect the truth "as seen clearly from the ground" and represent the ideal values that algorithms are striving to produce automatically.

Human judges carry out annotation on the basis of an *annotation protocol*, developed by the evaluators. The annotation protocol should clearly set out the **assumptions** that human judges should make when annotating the data. For example, the minimum length that a music segment must have in order to be considered a distinct segment or an inventory of the types of genres that can be used to label a piece of music.

The annotation protocol should not contradict the human judges' intuitions about what is the "right" way to annotate the data, but rather it should provide clarity for making decisions in ambiguous cases.

It is good practice to have several (e.g., three) judges to annotate all the data and then check for the consistency of their annotations (i.e., inter-annotator agreement). For some tasks, it is important to have judges with the right set of expertise (for example, the ability to distinguish the different instruments playing in the orchestra) and non-expert judges will not be able to achieve agreement. It is also good practice to check how the judgements of the annotators develop over time (i.e., intra-annotator agreement). If there is wide variability in either, the evaluation protocol should be refined. Including key examples can help increase the ability of the judges to consistently assess the borderline cases. If variability persists, then the evaluators should consider whether they are trying to evaluate a stable phenomenon. Another possibility is

that judgements follow a bimodal pattern and that underlying the problem are two different perspectives that can equally be considered "right".

Depending on the nature of the task, ***crowdsourcing techniques*** can be used to generate the ground-truth for a data set. Crowdsourcing is the process of "micro-outsourcing" small tasks to a general public crowd in return for a small payment. Popular commercial crowdsourcing platforms include Amazon Mechanical Turk and CrowdFlower.

When crowdsourcing is used for data set development, **careful attention must be paid to quality control of the annotations** created by the crowd. It is necessary to thoroughly pilot a crowdsourcing task before letting it run on the full data set. For the development of a highly effective task, one person month of researcher time must be planned.

Finally, in certain cases human annotators are not needed, but **automated techniques** can be used to approximate ground truth labelings. For example, within Task 3.3 of the project, the SDR (Signal-to-Distortion Ratio) can automatically be applied along with perceptually motivated measures. SDR approximates a human ground truth labeling, as it gives insight onto how humans perceive a stimuli. If such techniques are appropriate for the problem of focus, they are highly encouraged since they are the least expensive in terms of human and time resources.

Once the data has been evaluated, it is time to run the evaluation. Here, **two prototypes** or algorithms should be compared, one representing the **baseline** and the other being the **new algorithm** that is being evaluated. Evaluation takes place by applying an **evaluation metric** to compare the output of the algorithm to the ground-truth that has been produced by the human annotators. Evaluation produces a **score**, and also a **difference in score** that reflects the change in performance (hopefully an improvement) of the new system over the baseline.

After evaluation, it is important to take to further steps. First, ***statistical significance tests*** should be carried out in order to determine whether the improvement that has been achieved by a new algorithm versus the baseline is meaningful. Next, ***failure analysis*** should be carried out.

Failure analysis is the process of manually inspecting cases in the data in which the algorithms failed to perform as expected. It is a time consuming process, so it is important to choose the data that will be hand inspected wisely (for example, pick the most "suspicious" cases). Failure analysis has the goal of **discovering systematic problems** with the algorithm that can be corrected by developing the algorithm further. As such, it makes an important contribution to formative evaluation.

The following PHENICX tasks anticipate the use of ground-truth-based evaluation for the algorithms and protypes that they develop (cf. Deliverable 2.3 Technical Success Criteria, Section 3 page 6):

- Task 3.2: Multifaceted and musically meaningful analysis of audio streams and recordings
- Task 3.3: Multi-perspective audio processing techniques
- Task 3.4: Multimodal support for audio processing techniques
- Task 3.5: Web information extraction for different musical entities

- Task 4.1: Methods for extracting expression-related features from performance audio
- Task 4.2: Methods for recognising performer's and conductor's gestures.
- Task 4.3: Methods for synchronising recorded performances with scores or other performances (off-line), and for reliable live performance tracking
- Task 4.4: Methods for extracting expression-related features from a combination of recorded multimodal performance sources
- Task 5.5 Improving metadata-based matching of music items at different level of specificity

3.2 Output-based evaluation

For output-based evaluation to be most effective, it is necessary to have access to a large number of human judges whose availability matches the time schedule of the evaluation well.

The judges assess whether the output of the algorithm matches what they would expect. Here again, an annotation protocol is important, so that judges can be sure that they are all interpreting the assessment task in the same way.

In the past, ground-truth-based evaluation has been preferred over output-based evaluation. One reason is that complete annotation of the data set makes it possible to apply metrics such as recall (the proportion of the totality of all relevant existing material that the algorithm has returned). If only the output is judged, it is not possible to know what the system has failed to correctly identify. This shortcoming does not apply to many use scenarios in which users carry out tasks that are focused on precision. Another reason that ground-truth-based evaluation has been preferred is that output-based evaluation requires more person power. Each time a new algorithm is tested, new human judgements have to be carried out. It is possible to save the human judgements and, by output-based evaluation, ultimately arrive at something that approaches in usefulness a full-annotated data set. However, careful planning is required in advance such that judgements are not wasted.

However, currently, we are experiencing a new era of output-based evaluation because of the availability of crowdsourcing platforms mentioned above. Crowdsourcing is relatively inexpensive and can help to counter the practical challenges of collecting human judgments in the lab, which include human fatigue and the availability of space and equipment.

Within PHENICX, new opportunities exist because of the large number of users who are interested in music. Instead of using a commercial crowdsourcing platform, PHENICX is able to form at least a small music-oriented crowd by connecting with users via social media and asking for their feedback or to participate in evaluation.

The challenge of output-based evaluation is **reproducibility**. Each time new output is judged, it is important to ensure that the human judges are approaching the problem in the same way and that their judgements are stable, even in the face of changes in the composition of the judge-pool. As with ground-truth-based evaluation, statistical significance and failure analysis are important tools to ensure that the evaluation yields results that are meaningful and useful.

The following PHENICX tasks anticipate the use of output-based evaluation for the algorithms and protypes that they develop (cf. D2.3 Technical Success Criteria, Section 4 page 11).

- Task 5.2 Inferring community structures and socially established utility and meaning of music
- Task 5.3 Exploiting typical information needs and corresponding information and processing expectations
- Task 6.1: Visualisation of music pieces and their performances
- Task 6.2: Personalised multimodal information system
- Task 6.3: Acoustic rendering of augmented music performances

We finish by noting that **within PHENICX output-based-evaluation of algorithms and prototypes forms a continuum with use evaluation of Integrated System Prototypes**. Because output-based evaluation involves a large number of human judges, it is in effect, a user study that yields a highly quantified result measuring a very specific aspect of system performance. For example, Task 5.4 "Matching users at different levels of specificity" will be measured by investigating performance of the personalized prototype produced by Task 6.2. Here, it would be artificial to speak of a strict division between output-based evaluation and user studies.

4 THE DEVELOPMENT AND EVALUATION PROCESS FOR INTEGRATED PROTOTYPE SYSTEMS

In this chapter, a methodological outline is provided for the development and evaluation of integrated prototype systems. An overview of the design and development process is provided in Section 4.1, emphasizing how end-user feedback is used for the iterative development of integrated prototype systems. Section 4.2 contains a systematic description of methods that can be used as part of both the formative and the summative evaluation process. Section 4.3 contains an overview of user based performance indicators, which are based on the non-functional requirements and success criteria that have been formulated in Deliverable 2.3 "Technical Success Criteria".

4.1 Overview of the design and development process

In PHENICX, from a technical point of view the design and development process consists of:

1. The design and development of prototypes in the respective work packages.
2. Integration of prototypes.
3. Interface and interaction design for the Integrated System Prototypes (WP6).

Before, during, and after each of the sub processes **user-centered design activities** are planned. Depending on the phase of the design process, these activities are aimed at:

1 Formative evaluation: establishing user needs and user requirements (UR),
2 Formative evaluation: getting feedback on the user experience of integrated prototype systems (UE),
3 Summative evaluation: assessing the user-based success criteria (SC).

Establishing user needs and user requirements
A detailed planning will be made to align the user-centered design activities with the software development. Before the start of the development process, three focus groups will be run to understand the needs of the different target groups and to get feedback on the initial use cases. This feedback will be used to update the use cases. It will also be used to derive the user requirements.

User feedback on the user experience and underlying technology of the IPS's
At MS1, the first version of the integrated prototype systems will be ready. After each subsequent milestone, the systems will be evaluated with end-users. In between milestones other activities will be planned. These activities will be planned depending on the progress of the software development and the specific user-related questions that need to be resolved at that particular point.

The feedback from end-users will result in improvements to:

1. the individual prototypes that make up the integrated prototype systems (WP3-5),
2. the user interface of the integrated prototype systems (WP6).

The user-centered design techniques that we will use will elicit feedback from end-users with regard to:

1 *Technical performance:* How do users evaluate the resulting functionality that has been developed in WP3-5? What can be improved to make it more useful?
2 *User experience:* How do users evaluate the interaction mechanisms that have been evaluated in WP6? How can the user experience be improved?
3 *Technology acceptance:* to what extent do users accept the new technology as part of their concert experience? What changes can be made to alleviate the users' potential concerns?
4 *Usefulness and enrichment*: to what extent do the integrated prototype systems enrich the concert experience for users? In other words: to what extent are the non-functional requirements as defined in Deliverable 2.3 Technical success criteria addressed and how can this be improved?

Over the course of the design and development process the emphasis of the formative evaluation will move away **from technical performance and user experience towards usefulness and enrichment**: technological maturity and good user experience design are preconditions that need to be met before a user's attention can be drawn to the effectiveness and impact of the application.

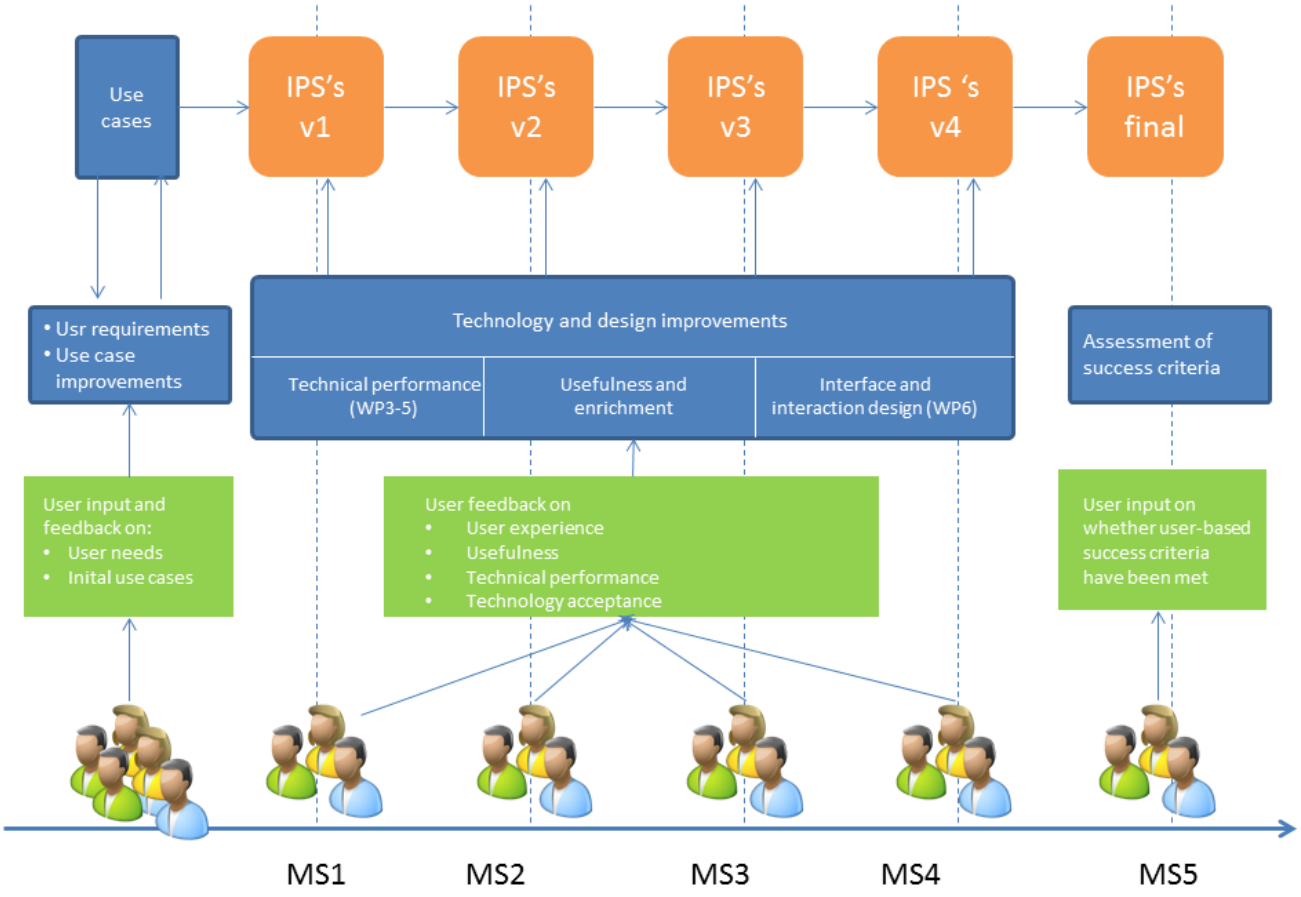

Figure 1 Relation between user-centered design activities and software development, sketching intended progress of Integrated Prototype Systems (IPS's) over the course of the different foreseen milestones (MS).

In Figure 1 a summary is provided of the relation between the software development process and user-centered design activities.

Assessment of success criteria
During the final evaluation we will assess whether the success criteria have been met. The evaluation of technical performance criteria has been addressed in the previous chapter. However, the non-functional requirements will also need to be operationalized into measurable success criteria.

4.2 Performance indicators for the summative evaluation

In Deliverable 2.3 "Technical success criteria", for each use case a preliminary list of non-functional requirements has been specified (Section 5, page 15). In the table below, we present a list of user-based performance indicators for each of the non-functional requirements. This list builds on, and substantially extends, the preliminary list.

Use case	Non-functional requirement	Performance indicators
Digital program notes	The application should feel like the reinvention of the music program booklet	Social influence Effort expectancy Perceived usefulness Added value of application over traditional program booklet User satisfaction
Virtual concert guide	In the on-site live case, this route should be minimalistic and respectful of the physical environment; In the off-site live case, the route should be adapted to the situation and thus offer more options of enrichment if desired by the user.	Social influence Effort expectancy, including readability (on-site) Performance expectancy Enrichment in relation to traditional concert guide User satisfaction
Overseeing music	The presentation of the scores - or variations of music visualisations on a digital screen need to be 'readable'. This implies separation of presentations for expert users and lay users.	Perceived added value of concurrent score visualization. Readability Salience of score visualizations Comprehension Effort expectancy Performance expectancy
Focusing attention and switching viewpoints	Zooming and edit functionalities need to be 'dummy proof'; Switching between point of	Effort expectancy Perceived usefulness

	view camera's has to be dummy proof.	
Comparing different performances	A graphical interface that envisions the differences between different versions.	Effort expectancy Performance expectancy Perceived added value compared to 'manual' comparisons
Capturing the moment	Capturing/adding a moment should be easy and feel natural.	Effort expectancy Performance expectancy Social influence Extent to which user interactions approximate natural behavior
Sharing the magic	Becoming a guide: users should get the feeling that they can become a guide within a musical piece.	Social influence Extent to which user interactions approximate natural behavior Affective responses
Joining the orchestra	The experience should not be a gimmick, but really connect with a serious music experience; Creating your style: a user should get the feeling that he is making a performance and not just accomplishing a desired task.	Extent to which user interactions approximate natural behavior Affective responses
Editorial support	UI design follows the desires of the system user (opposed to the needs of the end user).	Performance expectancy Effort expectancy Added value in relation to existing systems

As can be seen from the table, the success indicators can be clustered as follows:

- **Added value**: the added value of the developed applications in relation to their current day counterparts.
- **Techology acceptance**: using the unified theory of user acceptance and technology [Venkatesh et al. 2003], technology acceptance is measured with regard to effort expectancy, performance expectancy, and social influence.
- **Natural behavior**: the extent to which user interactions approximate natural behavior

Once the specifications of the integrated prototype systems—and, as a result the human-media interactions—have become clearer, we can start to define the user-based performance indicators more precisely.

4.3 Methods for user-based formative evaluation

In order to collect the data that is needed for both the formative evaluation and the summative evaluation, a large variety of methods is available. In this section, we present a selection of the methods that are useful within the PHENICX-project. Their applicability is specified according to the three types of user-centered design activities that have been specified in Section 4.1: establishing user needs an user requirements (UR), getting feedback on user experience design (UE), and assessment of success criteria (SC).

Method	Description	Applicability	Output
Logging	Logging functionality should be planned for and implemented in order to monitor user-system interactions. Implementing logging functionality however requires careful planning in terms of the data that should be collected and the output of the logging – to enable an efficient analysis. It can yield valuable insights with regard to the user interface, in particular with regard to the user's navigation through the apps. Interpretation of the data can be troublesome. Therefore, it should be combined with other methods	UE, SC	• Quantitative • Detailed event reports: which user did what on what page
Surveys	Surveys in different formats can be used to collect quantitative feedback, both to assess the success criteria and to provide designers with improvement suggestions. Online survey tools are available that can be embedded in applications, with a varying amount of obtrusiveness for the user (e.g. Kampyle).	UE, SC	• Primarily quantitative results • Assessment of the user-based success criteria • Integrated online survey tools can yield detailed feedback on the user experience.
Focus groups	Use interaction between 8-10 participants to collect feedback on design ideas, prototypes, and/or consensus about design decisions	UE, UR	• Feedback on design ideas, prototypes, visualizations
Experience sampling	Approach to collect longitudinal results by means of short surveys, triggered by time or certain events. Strength of the technique is the in-situ feedback: researchers get reliable results right after the behaviour of interest occurs. Surveys can both be paper-based, or based on the usage of any electronic device (such as smartphones and tablets).	UE, SC	• Qualitative and quantitative • Reliable in-situ feedback on the applications almost at the same time the user behaviour of interest occurs
Observation	Systematically observing users according to a pre-defined observation scheme. In PHENICX, observation can be used for the evaluation of the tablet apps. It can both be used to elicit feedback on the user experience and to assess technology acceptance in the concert hall and surrounding areas. If used for user experience evaluation, observation can be combined with *think aloud* in which participants are asked to verbalize every thought that comes to their mind.	UE, UR	• Qualitative • Detailed insight into user behaviour • *(if combined with think aloud)* detailed feedback on interface and interaction design aspects.
Critical	CIT is a systematic interviewing approach	UR	• Qualitative

incident technique	to have respondents recall particularly noteworthy episodes from their memory in relation to the object of interest. This approach can be useful to interview music professionals to learn about the computer support they would like to use.		• Rich contextual data about a user's current practices • *Please note reliability of recalling episodes from memory is an issue*
Online A/B testing	Online A/B-testing is an approach in which the performance of two designs are tested and compared concurrently. It is primarily used in online marketing to measure differences in conversion rates. In PHENICX, it can be used to compare different interface designs developed in WP6.	UE, SC	• Quantitative, mostly based on logging data

(method descriptions are derived from the Knowledge Centre of the European Network of Living Labs, *http://knowledgecentre.openlivinglabs.eu/*)

The precise set up of methods that are going to be used during the formative and summative evaluation depends on a significant number of factors, such as the course of the development process, the maturity of the integrated prototype systems at each of the milestones, and the availability of participants.

5 AN END-USER INVOLVEMENT STRATEGY

5.1 A user panel as an involvement strategy

In Chapter 2, we have addressed the importance of involving groups of end-users in the development of integrated prototype systems as well as the development and evaluation of prototypes. In the previous chapter, we have outlined the methods that can be used to get the desired input from end-users.

The important role of end-users in this project asks for an *involvement strategy* that will result in easy access to groups of end-users. This will allow us to quickly schedule and execute user-centered design activities that are well-aligned with the software development process.

Our aim is to have a panel of end-users that belong to one of the types of end-users we have distinguished in Deliverable 2.2 "Use cases document":

1 *Professionals: people who do not just wish to enjoy music, but will also use their concert experience in their own professional practice (musicians and music students).*

2 *The outsider, someone not involved with the genre, and not yet planning on getting confronted with the genre or going to concerts of it - while this could change when he would get intrigued by it.*

3 *The casual consumer, who passively knows the genre (e.g. by listening to it on the radio), without having deep knowledge of it. Does not mind going to concerts, but is not actively working on actually going there.*

4 *The heavy consumer, who actively knows the genre and all of its rituals. May be an (amateur) musician himself, and goes to concerts several times a year.*

Being a member of the panel means:

1. That the participants **signs a general informed consent form**, in which s/he is informed about—among others points—how his/her privacy is protected, how research data is treated, and what is expected from the participant.
2. That the participant **fills out a background survey** that asks for personal characteristics.
3. That the participant **allows the project to ask for his/her availability when a user study is planned**. Thus, participation is not obligatory for every session but is according to availability.

Such a panel requires an analysis of what the characteristics of different types of end-users are, where the users can be found, what channels they use, and what motivation to become a member of the panel we should appeal to. The results of this analysis can be found in the next section.

5.2 Different user groups and recruitment strategies

In this section, we describe the different target groups, their characteristics, and how they can be recruited. As discussed in Deliverable 2.2 "Use cases document", PHENICX distinguishes two types of users: end users (who are the people who will undergo the new digital concert experience) and system users (assisting with the realization of the new digital concert experience whose workflow is supported by PHENICX technology, e.g. editors, producers). It is important to note that system users are not targeted by the strategies set out in this section, rather they will be recruited via the informal network of the partners.

It should be noted that the characteristics below are currently strongly oriented towards and inspired by the context of the RCO consumer base, since this is the largest available concrete partner-provided consumer base available to the project. However, international equivalents to any Dutch references will actively be sought in the context of the ESMUC partner.

5.2.1 User profile: the casual consumer

The casual consumer passively knows the genre (e.g., by listening to it on the radio), without having deep knowledge of it. Does not mind attending concerts, but does not actively seek out opportunities to do so.

Properties:

1. People vary significantly in ages. Average age however is substantially lower than in the heavy users-group. Point of reference should be late thirties.
2. People have an open mind towards new technology (the early adoptors, or the early majority).
3. People enjoy classical music without having a strong intrinsic motivation to pursue that interest
4. People are only superficially familiar with the etiquette of classical concerts. As such, they do not mind if new elements are introduced that deviate from convention.

Recruitment:

Recruitment should aim for:

1. Participants from 20 to 45 years old (ad 1).
2. Participants that are experienced with smartphones and/or tablets (ad 2).
3. Participants who listen to classical music once in a while and who attend a performance at least once a year. A performance can be a concert, but also a festival or any other open-air performance. (ad 3).

Recruitment channels:

- Contacting radio channel, not only as an advertisement on the radio, but also as a news item on their websites.
- Distributing flyers during RCO-concert to those that belong to the aforementioned age group.
- Distributing flyers during festivals or other venues where open-air performances are being held (max. travel distance to Amsterdam: 30 mins.).
- Call to amateur orchestras, choirs, and so on (f.i. Krashna at the TU Delft).
- Finding Facebook pages of music-related organisations (including events). Posting messages on their Facebook wall.

- Recruitment via snowballing.

User incentive

- *Intrinsic:* be among the first to enjoy a new concert experience
- *Extrinsic:* (if possible) free tickets or other tokens of gratitude.

5.2.2 User profile: the heavy consumer

The heavy consumer actively knows the genre and all of its rituals. May be an (amateur) musician himself, and goes to concerts several times a year.

Properties:

- People attend classical concerts on average at least every two months or more.
- People are on average over 45 years of age.
- People on average belong to the early majority, late majority or the laggards: there is a large variety in technology adoption and acceptance.
- People value the rules that are part of the classical concert etiquette. People don't appreciate deviations from what they are used to with regard to the concerts.

Recruitment:
Recruitment should aim for:

- Variation in the user base with regard to technology acceptance (cfm. the first focus group).

Recruitment channels:

- Members of Next/Vrienden.
- Distributing flyers before, in the break of or after concerts.

Incentives

- *Intrinsic:* appeal to user's altruism and ask them to think and talk about their passion.
- *Extrinsic:* free tickets or another token of gratitude.

5.2.3 User profile: the outsider

The outsider is someone not involved with the genre, and not yet planning on getting confronted with the genre or going to concerts of it - while this could change when he would get intrigued by it.

Properties:

- Music lovers with a broad taste.
- People that read music magazines in order to stay up to date about musical developments and the latest album releases.
- They might be familiar with some pieces and might appreciate them, even though they don't recognize them as such. For instance: music used in movies, parts of classical pieces that are used in popular music, and so on.
- People are on average between 16 and 50 years of age.

- People on average belong to the early majority, late majority or the laggards: there is a large variety in technology adoption and acceptance.
- People are used to pop/rock concerts and may not be aware of the different etiquette at classical concerts.

Recruitment:
Recruitment should aim for:
- Variation in the user base with regard to technology acceptance (cfm. the first focus group.

Recruitment channels:
- readers of general music magazines like Oor;
- visitors of music festivals (possibly of other genres);
- contacts of contacts: recruiting outsiders from the social circles of existing consumers;
- UIT-markt (a cultural festival at the start of the new cultural year);
- TEDx events.

Incentives

- *Intrinsic:* Participants are among the first to learn about new ways of enjoying music.
- *Extrinsic:* Free tickets, cd's or vouchers. Please note that the gifts here should not be related to classical music (or on request only).

5.2.4 User profile: the professional

The professional is someone who does not just wish to enjoy music, but who will also use their concert experience in their own professional practice (musicians and music students).

Properties:
- Professionals are expert consumers of classical music.
- Professionals are eager to learn more about the pieces and about the players.
- The concert experience is not only valuable in itself, but is also helpful for their own work as professional musicians.

Recruitment:
Recruitment should aim for:
- Involvement of both students and professional musicians.

Recruitment channels:
- The student part of the professional group can be involved via ESMUC.

Incentives
- Professionals have a strong intrinsic motivation for the genre.

5.3 Background survey

A background survey will be developed and administered to new members of the panel. A preliminary list of topics for the survey contains the following topics:

1 Demographics (age, occupation, proximity of a concert hall);
2 Technology acceptance (based on [Venkatesh et al. 2003]);
3 Technology use;
4 Music taste;
5 Level of expertise and experience with classical music, as a music consumer and/or performer, including the frequency with which concerts are attended.

5.4 Planning

A **systematically growing user panel** is foreseen, aiming at 50 users for Milestone 2 (MS2) when the first version of the integrated prototype systems are due. In the meanwhile, smaller small-scale evaluations will be planned that will help to establish and grow the panel: after each evaluation users that have been recruited can be asked to become a member of the panel.

6 CONCLUSION

This deliverable has presented an initial description of the evaluation methodology to be used in PHENICX in the form of a set of guidelines. It has covered evaluation methodology for both the individual prototypes and the Integrated Prototype Systems.

Two types of methodology that will be used for the evaluation of algorithms and prototypes have been discussion. First, ground-truth-based evaluation, which uses a data set that has been annotated by human judges prior to carrying out the evaluation experiments, and, second, output-based-evaluation in which human judges inspect output after it has been produced. The advantages and disadvantages of each have been discussed.

The deliverable has also provided an overview of the design and development process, focusing on the relationship between user-centered design activities that are aimed at:

- establishing user needs and user requirements;
- getting feedback on the user experience, the perceived technical performance of the prototypes, and on usefulness and enrichment;
- assessment of the user based success criteria.

User-based performance indicators have been formulated for each of the use cases, based on the non-functional requirements that were defined in Deliverable 2.3 "Technical success criteria". Finally, methods were presented that can be used as part of the user-centered design activities that will yield the feedback that is necessary for the design and development process.

In order to have sufficient users from the various PHENICX target groups available for the formative and summative evaluation, an end-user involvement strategy has been presented in which a user panel is planned. A detailed analysis of the target groups has been provided to guide the participant recruitment process.

The results of this deliverable provide the project with methodological guidance for both the iterative development of the prototypes and the integrated prototype systems. As such, it will be used by the workpackages that develop and evaluate the prototypes (WP3-5), and the workpackages that integrate the prototypes into integrated prototype systems (WP6-7).

After the initial focus group with users who are heavy consumers, which took place M5, in September and October (M8-M9), two focus groups will be planned to establish the user needs from the other target groups. The results of the focus groups will lead to an updated set of use cases. Furthermore, they will also inform the development of the individual prototypes.

7 REFERENCES

7.1 Written references

[Venkatesh et al. 2003] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis and Fred D. DavisSource. User Acceptance of Information Technology: Toward a Unified View. MIS Quarterly, 27(3): 425-478, 2003.

7.2 Web references

Knowledge Centre of the European Network of Living Labs:
http://knowledgecentre.openlivinglabs.eu/ , accessed July 16, 2013.