Project Title:	**i-Treasures**: Intangible Treasures – Capturing the Intangible Cultural Heritage and Learning the Rare Know-How of Living Human Treasures
Contract No:	FP7-ICT-2011-9-600676
Instrument:	Large Scale Integrated Project (IP)
Thematic Priority:	ICT for access to cultural resources
Start of project:	1 February 2013
Duration:	48 months

Deliverable No: D7.1

Assessment Plan

Due date of deliverable:	31 January 2014
Actual submission date:	26 February 2014
Version:	Version 5 of D7.1
Main Authors:	Vasileios Charisis (AUTH), Stelios Hadjidimitriou (AUTH) and Leontios Hadjileontiadis (AUTH)

Project funded by the European Community under the 7th Framework Programme for Research and Technological Development.

Project ref. number	ICT-600676
Project title	i-Treasures - Intangible Treasures – Capturing the Intangible Cultural Heritage and Learning the Rare Know-How of Living Human Treasures

Deliverable title	Assessment Plan
Deliverable number	D7.1
Deliverable version	Version 5
Previous version(s)	1-4
Contractual date of delivery	31 January 2014
Actual date of delivery	26 February 2014
Deliverable filename	Deliverable_7_1_final.doc
Nature of deliverable	R
Dissemination level	PU
Number of pages	65
Workpackage	WP 7
Partner responsible	AUTH
Author(s)	Vasileios Charisis (AUTH), Stelios Hadjidimitriou (AUTH), Leontios Hadjileontiadis (AUTH), Nikos Grammalidis (CERTH), Kosmas Dimitropoulos (CERTH), Filareti Tsalakanidou (CERTH), Alexandros Kitsikidis (CERTH), Giannis Chantas (CERTH), Spiros Nikolopoulos (CERTH), Ioannis Kompatsiaris (CERTH), Marius Cotescu (ACAPELA), Athanasios Manitsaris (UOM), George Kourvoulis (UOM), Anastasios Katos (UOM), Eleni Katsouli (UOM), Alina Glushkova (UOM), Christina Volioti (UOM), Francesca Pozzi (ITD-CNR), Lise Crevier-Buchman (CNRS), Martine Adda-Decker (CNRS), Claire Pillot-Loiseau (CNRS), Angélique Amelot (CNRS), Samer K. Al Kork (UPMC), Bruce Denby (UPMC), Joëlle Tilmanne (UMONS)
Editor	Leontios Hadjileontiadis (AUTH)
EC Project Officer	Alina Senn

Abstract	This document describes the assessment plan, developed within Task 7.1, for the technical assessment and usability evaluation of the system's independent modules, the integrated platform and its functionalities.
Keywords	Technical Assessment, Usability Evaluation, Assessment Plan, Case Studies

Signatures

Written by	**Responsibility- Company**	**Date**
Vasileios Charisis, Stelios Hadjidimitriou, and Leontios Hadjileontiadis	Partner Researchers (AUTH)	10/2/2014
Verified by		
Leontios Hadjileontiadis	(AUTH)	17/2/2014
Leontios Hadjileontiadis	(AUTH)	17/2/2014
Approved by		
Nikos Grammalidis	Coordinator (CERTH)	25/2/2014
Yiannis (Ioannis) Kompatsiaris	Quality Manager (CERTH)	25/2/2014

Table of Contents

1. Executive Summary

Cultural expression is not limited to architecture, monuments or collections of artifacts. It also includes fragile intangible live expressions, which involve knowledge and skills. Such expressions include music, dance, singing, theatre, human skills and craftsmanship. These manifestations of human intelligence and creativeness constitute our Intangible Cultural Heritage (ICH). ICH is at the same time traditional, contemporary and living, because it does not only refer to inherited knowledge but also to the renewal of contemporary cultural expressions. It refers to the past, to the present, and, certainly to the future and is the mainspring of humanity's cultural diversity. The main objective of i-Treasures is to develop an open and extendable platform to provide access to ICH resources, enable knowledge exchange between researchers and contribute to the transmission of rare know-how from Living Human Treasures to apprentices.

This document was commissioned to present the assessment plan, developed within Task 7.1, for the assessment and evaluation of the system's independent modules, the integrated platform and its functionalities. The assessment and evaluation of i-Treasures project will be accomplished in two distinct phases. The first phase entails the laboratory testing of system modules and two cycles of case studies evaluation. The second phase comprises of the overall evaluation of the integrated platform.

The design of the assessment plan was based on two perspectives, the technical perspective and the user perspective, i.e., the performance of the technology used (technical performance assessment) and the evaluation of system's usability (usability evaluation). The technical performance assessment is imperative in order to assure optimization of the developed technological modalities, processing algorithms and interfaces. In this direction, assessment indices, organized in corresponding general categories, are introduced. These performance indices are divided into functional and non-functional. The functional assessment indices are based on user requirements and systems specifications (as described in deliverables D2.1 and D2.2), they are module specific and describe the functionality of the module. The non-functional criteria are common across the platform modules and use cases and relate to principal properties and characteristics of the system. The usability evaluation is also imperative since the i-Treasures platform aims towards knowledge exchange. The usability is going to be evaluated through case studies formulated for each sub-use case.

The overall evaluation of the integrated platform involves both assessment of major technical criteria and indices, as well as usability evaluation. For the overall evaluation to be realized, general performance indices (GPIs) will be generated for each sub-use case and use case. The GPIs will be estimated by fusing technical assessment and usability evaluation indices with the aid of fuzzy inference systems. Additionally, the GPI of the integrated platform will be determined that will reflect its overall quality and the degree to which the requirements of the potential users will have been met.

2. Introduction

2.1 Background

The aim of the i-Treasures project is to develop an open and extendable platform to provide access to intangible cultural heritage (ICH) and propose a novel strategic framework for the safeguarding and transmission of ICH by using novel multisensory technology for the creation of cultural content that has never been analyzed before. Such an ambitious and complex project requires a robust and independently verifiable performance assessment framework that will monitor progress towards objectives effectively and efficiently and will verify the compliance with the initially planned and intended impact.

The hitherto progress of i-Treasures project includes the successful completion of the first phase of work package 2 (WP2), Requirements and Specifications Identification. More specifically, within Task 2.2, all four use cases (rare traditional songs, rare dance interactions, traditional craftsmanship and contemporary music composition) were analysed and the first set of user requirements was specified. This set is contingent and suitable for further enhancement and will act as a reference guide. Moreover, the completion of Task 2.3 delivered the first detailed high-level reference architecture and specifications of the system and its components based on the definition of user and system requirements as well as on various technical, economical and legal constraints. The aforementioned information is meticulously described in documents *D2.1 First Report on User Requirements Identification and Analysis* and *D2.2 First Report on System Specification*, respectively, delivered so far.

These first reference user requirements and system specifications developed within WP2 form the stepping stones for the design of an efficient and realizable technical assessment and evaluation plan. The aim of WP7, namely Technical Assessment and Evaluation, is to formulate and apply an akin plan on i-Treasures platform. More particularly, the objectives of WP7 are as follows:

- To provide a methodological framework for assessing the performance of the integrated platform in terms of covering the user requirements and expectations.
- To carry out laboratory testing of the proposed system functionalities, in terms of technical and operational feasibility.
- To evaluate the proposed system, in terms of technical performance, user acceptance and cost-effectiveness.

The assessment - evaluation of the i-Treasures system aims at the assessment of the system functionalities from a technical point of view in terms of complying with the defined system specifications as well as from a user point of view in terms of complying with the user requirements and expectations. The development of the i-Treasures platform involves the assessment - evaluation of its functionalities from the design phase down to the demonstration phase, providing feedback to the development team about any detected shortcomings. Since the development approach is organized in two cycles, the assessment - evaluation for the first cycle (formative assessment-evaluation) will constitute the basis for design and development in the second cycle. Hereafter, the term ''assessment'' will refer to the assessment of the technical performance of the system modules and the platform as a whole. The assessment process will provide valuable feedback for identifying areas for technical improvement during the development phases. On the contrary, the term

"evaluation" will refer to a judgmental process that will gauge the quality of the system in terms of usability, user acceptance and satisfaction.

2.2 Aim of this Report

This document, named D7.1 Assessment Plan, is the first deliverable of WP7 of the i-Treasures project. The scope of this deliverable is to provide the assessment framework, developed within Task 7.1, for the assessment - evaluation of the i-Treasures system modules and sensor technologies, as well as, the i-Treasures platform as a whole. The user requirements and system specifications determined during the Tasks 2.2 and 2.3 and described in deliverables D2.1 and D2.2, respectively, are considered as a starting point in order to define appropriate assessment categories, objectives and measurable indices towards the construction of a detailed assessment plan.

2.3 Report Structure

The structure of this document is the following:

- Section 3 provides an overview of the assessment-evaluation methodology engaged as well as the organisation and scheduling of the assessment-evaluation process.
- Section 4 describes the first sub-framework of the general evaluation plan that is the laboratory testing of system modules (i.e., facial expression and modelling, body and gesture recognition, electroencephalography (EEG) analysis, vocal tract sensing and modelling, sound processing, text-to-song, data fusion, semantic analysis, educational platform and web platform interface) from a technical perspective.
- Section 5 presents the use cases evaluation sub-framework from the user's perspective that engages real users, both experts and learners, in order to perform usability evaluation. Moreover, section 5 introduces case studies for each use case scenario (i.e., rare traditional songs, rare dance interactions, traditional craftsmanship and contemporary music composition).
- Section 6 elaborates on the final assessment - evaluation of the i-Treasures integrated platform as a whole in order to assess its technical performance and validate it against the user requirements.
- At last, Section 7 is the Appendix that presents a technical assessment report template and a case study evaluation report template.

3. Overview of the Assessment - Evaluation Framework

In this Section an overview of the assessment – evaluation framework is presented by introducing the general categories that are taken under consideration, the rational, as well as the chronological organization of the assessment – evaluation process.

3.1 General Assessment - Evaluation Methodology

In general, the i-Treasures project aims at developing a technology-based system that will capture aspects of intangible cultural heritage and make them accessible to a wide range of users for both informative and educational purposes. To this end, the system incorporates different technology modalities that will be recruited to accomplish an efficient, effective and satisfactory conveyance of the intended information to users, both learners and experts in the related field. To foster the optimization of the aforementioned characteristics of the system, i.e., efficiency, effectiveness and satisfaction, an assessment - evaluation process has to be implemented during the development and testing phases of the system. Figure 3-1 offers an overview of the expected evolution of the development and assessment – evaluation processes of i-Treasures.

The methodology adopted here focuses on two perspectives, i.e., the performance assessment of the technology used (technical performance assessment) and the usability evaluation of the system (usability evaluation). As far as the technical performance is concerned, scientific expertise is required in order for the related categories to be properly assessed, while usability evaluation mandates valuable feedback from the users' perspective. In particular:

- *Technical Performance Assessment:*

  Technical performance assessment is critical in order for an optimised implementation of the technological modalities, such as sensors, processing algorithms and interfaces, to be achieved. As these modalities require scientific knowledge and expertise it relies mainly on the researchers involved in the project to perform the assessment. To this end, assessment categories and corresponding indices are introduced which are based on specifications of the system architecture (deliverable D2.2 *First Report on System Specification),* as well as the functional and the non-functional requirements of the system (deliverable D2.1 *First Report on User Requirements Identification and Analysis*) identified in work package WP2 *Requirements and Specifications Identification.*

  The categories and indices based on non-functional requirements are common for all system modules and use cases and include system properties that are fundamental, e.g., cost and common technical framework characteristics like scalability and interoperability.

  Additionally, specific technical performance categories and indices are introduced for each system module due to the different characteristics of each one: a) facial expression and modelling, b) body and gesture recognition, c) electroencephalography analysis, d) vocal tract sensing, e) sound processing and f) text-to-song module. Moreover, general categories and indices concerning data fusion and semantic analysis are introduced to assess their performance in the four defined use cases, i.e., rare traditional singing, rare dance interactions, traditional craftsmanship, and contemporary music composition. In the same vein, specific assessment categories are also introduced for the system interface, i.e., the 3D visualisation module, the learning management system, and the web platform.

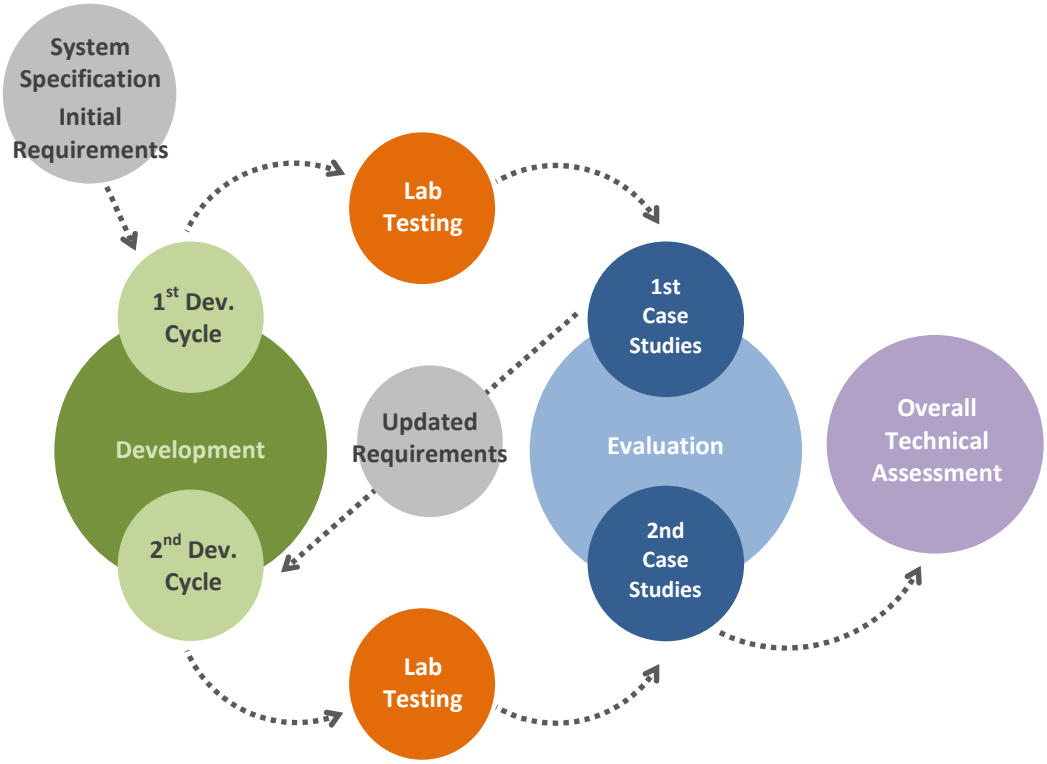

Figure 3-1. Interwoven evolution of the development and assessment – evaluation processes of i-Treasures.

Generally, the technical performance assessment of i-Treasures will be an ongoing activity covering both the first and second development cycles of the system and it will mainly take place during the laboratory testing (Task 2.1 Laboratory Testing of Modules). Valuable feedback is expected after the completion of the first phase of use cases evaluation.

- *Usability Evaluation:*

  Usability is a critical characteristic of a system like the i-Treasures platform that aims at conveying knowledge for both informative and educational purposes to a wide range of users. In order to evaluate the usability of the system, a series of use case evaluation tests (Task 7.3 *Case Studies Evaluation*) will be conducted by organizing case studies for each intended use.

  The objectives of the case studies and their design are partly based on the requirements that mandate user feedback, identified from the deliverable D2.1 of work package WP2. Real users, both experts (12-33 experts) and learners (60-180 learners) will participate in the tests that will include capturing, educational, and information retrieval scenarios. In particular, for each use case, the following case studies are expected to be carried out:

  - Rare Traditional Songs: Case Study 1: Byzantine Music; Case Study 2: Canto a Tenore; Case Study 3: Cantu a Paghjella; Case Study 4: Human Beat Box
  - Rare Dance Interactions: Case Study 1: Traditional Tsamiko/Calus Dancing; Case Study 2: Traditional Walloon Dancing; Case Study 3: Contemporary Dancing;

- Traditional Craftsmanship: Case Study 1: Traditional Craftsmanship
- Contemporary Music Composition: Case Study 1: Contemporary Music Composition

Considering the case studies, usability data will be acquired using both traditional techniques (e.g. questionnaires) and automated tools (software tools for usability data collection), and user satisfaction indices will be constructed. Data will be analysed and significant factors and their interactions will be derived for examining the extent the identified user requirements were met, and accordingly, updating decisions to be taken (deliverable D2.3 *Second Report on User Requirements Identification and Analysis*).

Case studies will take place in two phases. The first phase is intended to evaluate the developed system of the first cycle and provide feedback (e.g. updated user requirements) for the second development cycle. The second phase will be carried out during the second development cycle in order to reevaluate the altered system and it will include updated – based on the revised user requirements – versions of the first phase case studies.

After the completion of the assessment – evaluation process that targets the two development cycles of the i-Treasures platform, the system will further be tested during a final Technical Assessment of the System process (Task 7.4). The aim of this phase will be to assess the overall technical performance of the system and validate it against the final user requirements (deliverable D2.5 *Final Report on User Requirements Identification and Analysis*). To this end, appropriate general performance indices are introduced based on both pure technical assessment indices, as well as usability evaluation indices, through a fuzzification process. These general performance indices will be calculated by employing representative tests of the system for each intended use case. For the sake of comparison, pure reference cases and their corresponding general performance indices will be determined based on the assessment – evaluation evidence obtained during the development process of the system, especially after the second phase of case studies. The overall technical assessment of the platform is expected to further improve its performance and contribute to its optimization.

3.2 Organisation and Scheduling of Assessment - Evaluation

The assessment - evaluation process of the i-Treasures platform is planned in two distinct phases with respect to time. The first phase, spanning over the two development cycles of the system, will take place from month 13 to month 39 and it includes the technical performance assessment of the system and its modules as well as the case studies evaluation. The second phase starts immediately after the first, i.e., in month 40, and lasts until the end of the project. During that period, the overall technical assessment of the developed system will be carried out by measuring the general performance indices and testing them against the already-defined reference cases. For an overview, Figure 3-2 illustrates the chronological organization of the assessment – evaluation process of the platform.

In the first phase, laboratory testing of modules (Task 7.2) will focus on the technical performance assessment and it will last from month 13 to month 36. It will comprise three assessment subcategories, i.e., the technical performance assessment of system modules, the technical performance assessment of data fusion and semantic analysis, and the assessment of the system interfaces. The technical performance assessment of system modules starts at month 13, near the end of WP3. *ICH Capture and Analysis* (months 3 - 14) when the first development cycle of the modules will be almost completed, and lasts until month 36. The assessment of the

technical performance of data fusion - semantic analysis and of the system interfaces begins at month 18, near the first halves of work packages WP4 *Data Fusion and Semantic Analysis* and WP5 *The Integrated Platform for Research and Education*, and it also ends at month 36.

Additionally, during the first phase, the use cases evaluation (Task 7.3) will take place from month 21 to month 39. Case studies are planned to be conducted in two phases. The first phase of case studies will last from month 22 to month 24, providing initial usability evidence, while the second phase will be carried out between month 37 and month 39, after the analysis of users' feedback and the completion of the technical performance assessment process. In particular, case studies per use case are expected to proceed according to the following proposed time schedule:

Rare Traditional Songs

- **Case Study 1 (UOM): Byzantine Music**

  Design/Preparation/Planning: month 22

  Data Collection: months 22-23 (First phase), month 37-38 (Second Phase)

  Data Analysis: months 23-24 (First phase), months 38-39 (Second Phase)

- **Case Study 2 (CNR-ITD): Canto a Tenore**

  Design/Preparation/Planning: month 22

  Data Collection: months 22-23 (First phase), month 37-38 (Second Phase)

  Data Analysis: months 23-24 (First phase), months 38-39 (Second Phase)

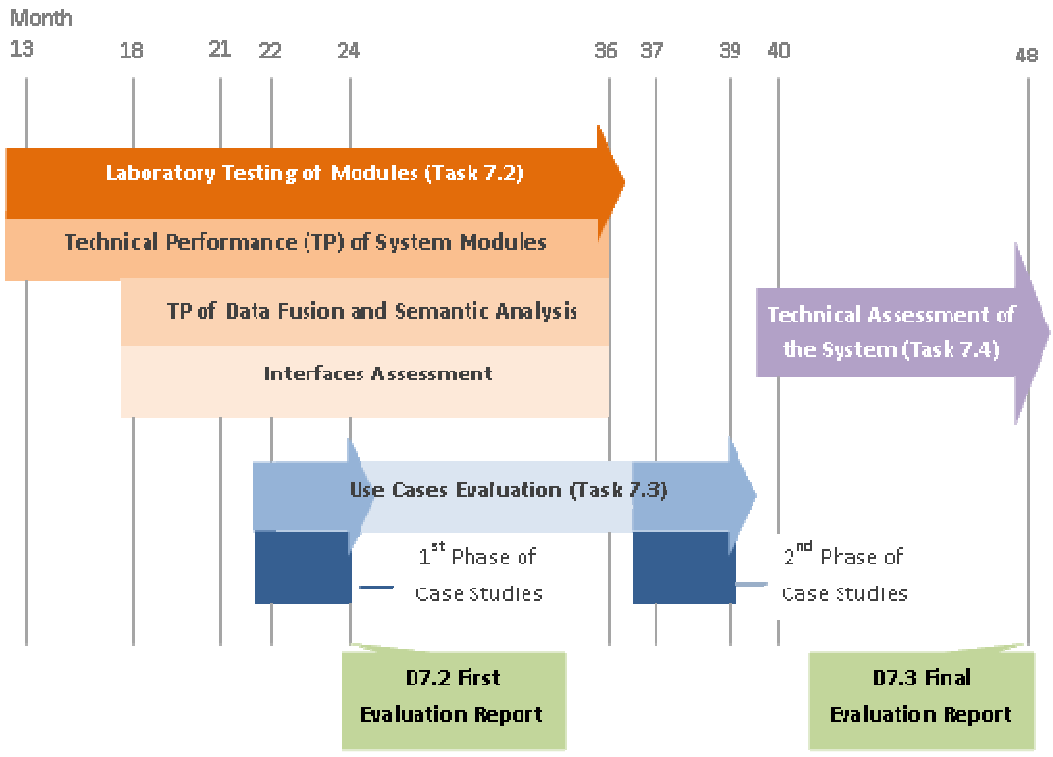

Figure 3-2. Organization of the assessment – evaluation process of i-Treasures.

- **Case Study 3 (CNRS): Cantu in Paghjella**
  Design/Preparation/Planning: month 22
  Data Collection: months 22-23 (First phase), month 37-38 (Second Phase)
  Data Analysis: months 23-24 (First phase), months 38-39 (Second Phase)
- **Case Study 4 (CNRS): Human BeatBox**
  Design/Preparation/Planning: month 22
  Data Collection: months 22-23 (First phase), month 37-38 (Second Phase)
  Data Analysis: months 23-24 (First phase), months 38-39 (Second Phase)

Rare Dance Interactions

- **Case Study 1 (CERTH): Traditional Tsamiko/Calus dancing**
  Design/Preparation/Planning: month 22
  Data Collection: months 22-23 (First phase), month 37 (Second Phase)
  Data Analysis: months 23-24 (First phase), months 38-39 (Second Phase)
- **Case Study 2 (UMONS): Traditional Walloon dancing**
  Design/Preparation/Planning: month 22
  Data Collection: month 22-23 (First phase), months 37-38 (Second Phase)
  Data Analysis: month 23-24 (First phase), month 38-39 (Second Phase)
- **Case Study 3 (UMONS): Contemporary dancing**
  Design/Preparation/Planning: month 22
  Data Collection: month 22-23 (First phase), months 37-38 (Second Phase)
  Data Analysis: month 23-24 (First phase), month 38-39 (Second Phase)

Traditional Craftsmanship

- **Case Study 1 (CERTH, ARMINES, UOM, TT): Traditional Craftsmanship**
  Design/Preparation/Planning: month 22
  Data Collection: month 22-23 (First phase), months 37-38 (Second Phase)
  Data Analysis: month 23-24 (First phase), month 39 (Second Phase)

Contemporary Music Composition

- **Case Study 1 (UOM): Contemporary Music Composition**
  Design/Preparation/Planning: month 22
  Data Collection: month 22-23 (First phase), months 37-38 (Second Phase)
  Data Analysis: month 23-24 (First phase), month 39 (Second Phase)

At the end of the case studies first phase, the First Evaluation Report (deliverable D7.2) will be delivered (month 24).

After the completion of both the technical performance assessment and the usability evaluation processes which will cover the two development cycles of the system, the integrated platform incorporating alterations based on the acquired feedback will be further assessed during the last nine months of the project (months 40-48). Chronologically, during that period of overall technical assessment (Task 7.4), reference cases per use case scenario will be established, representative tests for all use cases will be conducted during the system demonstration, and the general performance indices derived from the reference cases and the tests will be analysed and compared. The overall technical assessment phase will overlap with the demonstration of the i-Treasures platform as planned in work package WP6 *System Demonstration*, and at its end (month 48) the Final Evaluation Report (deliverable D7.3) will be delivered.

4. Technical Assessment Categories and Indices

This section presents the assessment criteria and indices that will be used for the technical assessment of the platform and its individual modules. Indices are divided into two main categories. The first category contains indices that are based on non-functional requirements and are common across the platform regardless of the use cases. The second category contains indices that are determined for each module separately as well as for the data fusion and semantic analysis, web platform, learning management system and 3D visualization entities. In Appendix 7.1 the template of the technical assessment report is provided. The design of this template is partly based on the principles specified in *IEEE 829-1998*, also known as the *829 Standard for Software Test Documentation* [1] and adapted to the i-Treasures platform specifications and characteristics.

4.1 Assessment Categories and Indices based on Non Functional Requirements

In this section, general assessment categories and indices are introduced that are based on the non-functional requirements identified in the deliverable D2.1 *First Report on User Requirements Identification and Analysis* of work package WP2 *Requirements and Specifications Identification*. The term *non-functional* refers to the requirements that are not critical for the system to function as desired; however, they are of significant importance to the usability, sustainability and maintainability of its final release. Criteria reported here conform to the *ISO/IEC 25010* standard for the quality of systems and software [2] and will be taken under consideration during both the first and second development cycles of the i-Treasures platform as well as during the overall assessment of the final system.

Nine general assessment categories are presented in the following tables along with corresponding lists of desired qualities – in the form of questions – that will be evaluated. In all cases, there is a common ordinal assessment index (fulfillment level (FL)) in the form of a discrete five-level scale (1 - 5), with 1 denoting the absence of the quality and 5 denoting the theorized maximum amount of fulfillment. However the impact of each quality on the functionality of the platform varies, and for this reason, an appropriate weight will be assigned to each of the criteria in the estimation of general performance indices during the overall assessment of the platform phase. In particular, using factor analysis, each of the nine constructs will be aggregated to produce nine first-order constructs, and by further aggregating these nine constructs, a single overall second-order construct will be developed. The various weights (factor loadings) will be used to evaluate the contribution of each construct to the final result.

1. Cost Optimality / Energy efficiency		
Quality	**FL (1-5)**	**Comments (if applicable)**
1.1) Is the infrastructure (e.g. servers, ups units, broadband services etc.) of the i-Treasures platform cost- and energy-efficient?		*Report the cost of the expected infrastructure*
1.2) Is the maintenance of the infrastructure cost-efficient?		*Report the maintenance costs*
1.3) Is it affordable for the single user – institution to acquire the necessary hardware equipment for		*Report the cost of the necessary sensors – recording devices for each use case*

the i-Treasures application (e.g. sensors, recording devices)?		
1.4) Does the user need to install on-fee third party applications for the system to function properly?		*Report the cost of the third-party applications*
1.5) Does the proper functioning of the i-Treasures platform requires updated hardware purchase (e.g. personal computer with new processor, amount of RAM etc.)?		*Report the minimum system requirements*
1.6) Will future versions of i-Treasures require the purchase of additional or upgraded hardware?		

Table. 4-1 Non-functional criteria referring to cost and energy efficiency

2. Accessibility / Usability		
Quality	**FL (1-5)**	**Comments (if applicable)**
2.1) Is the content of the i-Treasures presented in different forms (auditory, text, visual elements)?		
2.2) Is the web-platform content rendered properly across different browsers?		*Supported Browsers*
2.3) Is the i-Treasures web platform responsive?		
2.4) Are web-pages accessible even when newer web technologies are not available (e.g. JavaScript, CSS3 stylesheets)?		
2.5) Does the i-Treasures platform support different languages?		*Languages supported*
2.6) Does the platform provide clear navigation mechanisms?		
2.7) Does the platform support simple and advanced search?		
2.8) Is the content presentation consistent across pages?		
2.9) Does the platform provide feedback on connected hardware configuration?		
2.10) Are response times for task accomplishment (e.g. application loading, refresh time, search results etc.) proper and adequate?		

Table 4-2. Non-functional criteria referring to accessibility and usability

3. Documentation / Support		
Quality	**FL (1-5)**	**Comments (if applicable)**
3.1) Is there a complete, accurate and clear documentation accompanying the i-Treasures platform?		
3.2) Is the documentation partitioned into sections for users, developers and administrators?		
3.3) Does the assimilation of the documentation require background or expertise from the users?		
3.4) If yes, are there links to supporting information resources?		
3.5) Is there a "getting started" or a "how to" guide for different use cases?		
3.6) Is the documentation available in the i-Treasures website?		
3.7) Is there a web page (e.g. FAQs), a forum or e-mail lists for additional support?		

Table 4-3. Non-functional criteria referring to documentation and support

4. Interoperability / Portability		
Quality	**FL (1-5)**	**Comments (if applicable)**
4.1) Are the platform components/modules compatible with third party services?		
4.2) Does the platform function on different and commonly available operating systems (OSs)?		*Supported OSs*
4.3) Is the downloadable content compatible with the different OSs?		
4.4) Besides personal computers, is the platform compatible with tablets or other mobile devices?		

Table 4-4. Non-functional criteria referring to interoperability and portability

5. Extensibility / Scalability		
Quality	**FL (1-5)**	**Comments (if applicable)**
5.1) Are the core/custom components of the i-Treasures platform modular?		
5.2) Is backward compatibility taken into account during upgrades?		

5.3) Is the source code well-structured and according to coding standards?		
5.4) Can the system be remotely managed?		
5.5) Does the servers deployment allow for horizontal scalability?		

Table 4-5. Non-functional criteria referring to extensibility and portability

6. Auditing		
Quality	**FL (1-5)**	**Comments (if applicable)**
6.1) Does the i-Treasures platform provide reports on user activity (per application) and general user behavior?		
6.2) Does the platform support auditing of failed login attempts in order to detect brute force attacks?		
6.3) Is the auditing system centralized and secured?		

Table 4-6. Non-functional criteria referring to auditing

7. Security / Privacy		
Quality	**FL (1-5)**	**Comments (if applicable)**
7.1) Does the i-Treasures platform have an authentication system?		
7.2) Does the i-Treasures database support role-based access control based on user privileges?		
7.3) Is user data transferred and stored securely (e.g. use of encryption algorithms, HTTPS etc.)?		
7.4) Is user data available and to whom?		
7.5) Does the system user authentication take measures in cases of misuse (lost or stolen passwords, account locks)?		

Table 4-7. Non-functional criteria referring to security and privacy

8. Fault tolerance / Recoverability		
Quality	**FL (1-5)**	**Comments (if applicable)**
8.1) Is there a software failure monitoring procedure (e.g. error logs)?		

8.2) Are hardware maintenance and software upgrades planned in a systematic way?		
8.3) Does the system include regular backups of system components and images?		
8.4) Can system backup images be restored on different hardware?		
8.5) Does the architecture of the platform allow for autonomous functioning of services when a module fails and is being disabled?		
8.6) Is the platform functional in cases when external web dependencies (e.g. Europeana platform) go offline?		

Table 4-8. Non-functional criteria referring to fault tolerance and recoverability

9. Licensing / Copyright		
Quality	**FL (1-5)**	**Comments (if applicable)**
9.1) Has an appropriate license been adopted?		
9.2) Is the type of license clearly stated in all platform aspects (e.g. website, source codes)?		
9.3) Do platform components include a copyright statement?		
9.4) Is it clearly stated who funded the project, developed the platform and owns the copyright?		
9.5) Does the platform have a trademark, that doesn't violate other trademarks?		

Table 4-9. Non-functional criteria referring to licensing and copyright

Entries of the preceding tables, depending on the development phase of the i-Treasures platform, should be filled and included in both the first and final evaluation reports (deliverables D7.2 and D.7.3, respectively).

4.2 Technical Performance Categories and Indices of ICH Capture and Analysis

This section presents the technical assessment indices organised in general categories for each module separately. Every assessment index is accompanied by:

- a short description explaining which quality/feature is measured/assessed, the data type (Numerical, Qualitative, Continuous, Binary, Discrete, Ordinal etc.),
- the values that the index may acquire,

- the codes of the user requirements that might be examined by the specific index (if applicable), as defined in *D2.1 First Report on User Requirements Identification,* and
- the suggested assessment phase (laboratory testing - LT, case study evaluation - CS, overall assessment - OA) that the index will be used. However, this declaration is not binding. Every assessment index can and may be used in every assessment phase if the circumstances require so.

4.2.1 Facial Expression and Modeling

Table 4-10 presents critical performance indices that will be used for the technical assessment of facial expression and modeling module.

Module	**Facial Expression and Modeling**					
Assessment Category	**Assessment Indices**				**Requirements Examined**	**Assessment Phase**
	ID	**Description**	**Type**	**Values**		
Facial feature tracking accuracy	Local Error	Mean facial feature localization error in mm and/or pixels. To measure the performance of the face tracking algorithm, we compare the estimated feature positions against their real (ground-truth) positions.	Numerical	0-Inf	R3.H	LT
Facial Action Unit (AU) recognition accuracy	Conf MatAU	Confusion matrix: The element *(i,j)* of the confusion matrix represents the percentage of instances that AU *i* (ground truth) was classified as AU *j* (result of classifier).	Matrix with numerical values	0-1	R3.H	LT
	AccAU	Classification accuracy: ratio of correctly predicted AUs over the total number of predictions.	Numerical	0-1	R3.H	LT
Basic emotion recognition accuracy	Conf MatEmo	Confusion matrix: The element *(i,j)* of the confusion matrix represents the percentage of instances that emotion *i* (ground truth) was classified as emotion *j* (result of classifier).	Matrix with numerical values	0-1	R3.H	LT
	AccEmo	Classification accuracy: ratio of	Numerical	0-1	R3.H	LT

		correctly predicted emotions over the total number of predictions.				

Table 4-10. Assessment indices for Facial Expression and Modeling module

Confusion matrix: a specific table layout that allows visualization of the performance of a machine learning algorithm. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class (ground truth). The element *(i,j)* of the confusion matrix represents the ratio of instances that a sample from class *i* was classified as class *j* over the total number of instances of class *i*. The matrix makes it easy to see if the system is confusing two classes (i.e. commonly mislabelling one as another).

For the assessment of the module, the following test data sets will be used:

- Database of image sequences recorded by CERTH-ITI. The database will be comprised of sequences of 3D and corresponding 2D images showing human subjects mimicking basic facial expressions and performing a subset of the action units of the FACS system. This data set will be used for the laboratory testing of this module (1st assessment phase).
- i-Treasures data recordings. In the context of Task 3.6 “Data collection”, a set of 2D and 3D image sequences showing experts (composers and/or singers) performing their art will be recorded and will be used for algorithm assessment. This data set will be used for the laboratory testing of this module (1st assessment phase).
- i-Treasures demonstration recordings. In the context of WP6 “System demonstration”, a set of 2D and 3D image sequences showing experts (composers and/or singers) performing their art in real demonstrations will be recorded. This data set will be used in the second cycle of the system assessment (2nd assessment phase).

4.2.2 Body and Gesture Recognition

In the context of i-Treasures project, the objective of Human Body Motion and Gesture Recognition is based on the following steps:

1. Mod: Development and estimation of all the dynamic and simultaneous relationships that describe the body movements and/or gestures, where motion activities (e.g., dance steps or pottery making steps, gestures) have been annotated by experts.

2. Sim: Simulation (i.e., dynamic solution) of the system of equations estimated in step 1, and investigation of its predictability power. The simulated solution constitutes the base line for comparisons.

3. Sem: Sensitivity analysis of the dynamically simulated system of equations in step 2, investigating its behaviour for disturbances applied to various exogenous variables (e.g., initial dimensions) at specific levels.

4. Per: Comparison of data recorded in real demonstrations with the base line simulated data produced in step 2.

The table below presents the major criteria that will be used for the technical assessment of the body and gesture recognition.

Module	Body and Gesture Recognition					
Assessment Category	**Assessment Indices**				**Requirements Examined**	**Assessment Phase**
	ID	**Description**	**Type**	**Values**		
Quality of Human Body Motion and Gesture Recognition structure	Mod	Quality of each estimated equation describing the structure of the Human Body Motion and Gesture	Quantitative / Qualitative, using statistical significances	• Low • Medium • High	R3.C, R3.D	LT
Human Body Motion and Gesture recognition Accuracy	Sim	Accuracy of each dynamically simulated endogenous variable of the Human Body and Gesture Recognition system	Comparison between model simulated data and real data, Numerical, Continuous	• Theil: 0-1 (close to zero) • Bias proportion: 0-1 (close to zero) • Variance proportion: 0-1 (close to zero) • Covariate proportion: 0-1 (close to one)	R3.C, R3.D	LT
Sensitiveness of the Human Body Motion and Gesture recognition	Sen	Sensitivity Analysis using the dynamically simulated system of Human Body Motion and Gesture equations	Computation of multipliers with respect to disturbances at various levels, Numerical, Continuous	• Percentages (positive / negative) • Actual differences (positive / negative)	R3.C, R3.D	LT
Performance identification accuracy	Per	Comparison between base model simulated data (base line) and data recorded in real demonstrations	Numerical, Continuous	• Theil: 0-1 (close to zero) • Bias proportion: 0-1 (close to zero) • Variance proportion: 0-1 (close to	R3.C, R3.D	CS, OA

				zero) • Covariate proportion: 0-1 (close to one)		

Table 4-11. Assessment indices for Body and Gesture Recognition module

For the assessment of the module, the following test data sets will be used:

- Database of test data recorded from multiple depth sensors (e.g. Kinect) or from inertial or optical motion capture suits. The database will consist of fused skeleton data of recorded dancers or pottery makers, where motion activities (e.g. dance steps or pottery making steps) have been annotated by experts. The same applies for gestures. These data sets will be used for the laboratory testing of this module (1st assessment phase).
- i-Treasures data recordings. In the context of Task 3.6 “Data collection”, a set of fused skeleton data of dancers or pottery makers will be recorded. Motion activities (e.g. dance steps or pottery making steps) will be annotated by experts and will be used for algorithm assessment. A similar procedure will take place for gesture capturing. These data sets will be used for the laboratory testing of this module (1st assessment phase).
- i-Treasures demonstration recordings. In the context of WP6 “System demonstration”, fused skeleton data of dancers or pottery makers will be recorded in real demonstrations. In the same way, gestures will be also recorded. These data sets will be used in the second cycle of the system assessment (2nd assessment phase).

4.2.3 Electroencephalography Analysis

In the context of the i-Treasures project, the objective of electroencephalography analysis is to recognize different affective states of the user based on his/her brain activity in real-time. To this end, electroencephalogram (EEG) signals will be acquired, fed to a processing – classification algorithm, and mapped to an affective label or a quartile of the valence – arousal plane, i.e., a two-dimensional space for emotion characterization. Signal acquisition is planned to be conducted using the Emotiv EPOC EEG acquisition device and the succeeding algorithmic processing and mapping will function at refresh rates that verge on real-time behavior. Thus, the technical assessment of EEG analysis will focus on the quality of acquired signals, the efficiency of affective sate detection and the computational time by measuring the indices provided in the table below.

Module	**Electroencephalography Analysis**					
Assessment Category	**Assessment Indices**				**Requirements Examined**	**Assessment Phase**
	ID	**Description**	**Type**	**Values**		
Quality of EEG recording	EEG Qual	The quality of each EEG signal (per channel) as recorded by EPOC device	Qualitative	• Low • Medium • High	-	LT
Affective Sate	Num OfSt	The number of different affective	Numerical, Discrete	2 - 6	-	LT

Detection Efficiency	ates	states recognized.				
	Acc	Classification Accuracy: Ratio of correctly predicted states over the total number of predictions	Numerical, Continuous	0 - 1	-	LT
	Conf Mat	Confusion Matrix: Table that visualises classification performance. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class	Numerical, Continuous	0 - 1	-	LT
Computational Time	DetTime	Amount of time (in milliseconds) required to perform state detection	Numerical, Continuous	>0	-	LT

Table 4-12. Assessment indices for Encephalography Analysis module

EEG datasets, which will serve as ground truth, are expected to be acquired from participants through a series of experiments in a laboratory environment. As emotion evocation media, images and audiovisual content will be used. In addition, participants will provide self-reported assessment of induced affective states, through appropriate questionnaires, that will serve to define the target classes to which the EEG epochs will be categorized.

4.2.4 Vocal Tract Sensing and Modeling

Ultrasound images analysis is used to model tongue movements inside the mouth. The quality of the recognition of tongue movements depends on the quality of ultrasound images. A qualitative assessment index is defined. Tongue movements can be described by tongue surfaces that can be extracted from contour points from ultrasound pictures. Two assessment indices are used: one to check the learning performances and one to validate the final conversion into pixel coordinates. All of these indices focus on the quality of tongue movement detection. In addition, computational time must be taken into account and compared to acquisition rate to meet real time constraints. Regarding tongue contour extraction, a manual input can be used as ground truth for small datasets. However, since hand-labeling is time consuming, an automatic (but not real-time) tongue contour detection algorithm can be used instead as ground truth.

Module	**Vocal Tract Sensing and Modeling**					
Assessment Category	**Assessment Indices**				**Requirements Examined**	**Assessment Phase**
	ID	**Description**	**Type**	**Values**		
Quality of US pictures	USquality	Quality of ultrasound images (noise)	Qualitative	• Poor • Medium	-	LT

				• High		
Tongue contour extraction	Validation error	Squared difference between original ultrassound image and reconstructed ultrasound image	Numerical, Continuous	0 - 1	-	LT
Tongue contour extraction	MSD	Mean Square of Distance between manual contour used as ground truth and automatically extracted contour	Numerical, Continuous	0 - 5 (usually < 5)	-	LT
Computational Time	elapsedTime	Time required to perform a pass of the data through the network	Numerical, Continuous	> 0	-	LT,CS, OA

Table 4-13. Assessment indices for Sound Processing module

4.2.5 Sound Processing

In the context of the i-Treasures project, the objective of sound processing is to analyze, detect and recognize vocalizations that are part of the intangible cultural heritage. Amongst others: human beatboxing, corsican, sardinian, and byzantine singing.

More specifically, we focus for the moment on human beatboxing. A new high-quality beatbox database was recorded, consisting of beatbox sounds produced by a single male beatboxer. The database contains 4 sets: individual drums sounds, rhythms, instruments and freestyle. The audio was captured in a soundproof room at a sampling rate of 48 kHz. The audio signals are then analyzed using various algorithms (see table here below).

Module	**Sound Processing**					
Assessment Category	**Assessment Indices**				**Requirements Examined**	**Assessment Phase**
	ID	**Description**	**Type**	**Values**		
Individual drums sounds recognition Accuracy	Conf_Mat	Confusion matrix	Matrix with numerical values	0 – 1	R1A	LT
	Acc_Reco	Recognition accuracy	Numerical, Percentage	0 – 100 %	R1A	LT
	Acc_Lat	Latency-aware F-score	Numerical, continuous	0 – 1	R1A	LT
Instruments recognition Accuracy	Conf_Mat	Confusion matrix	Matrix with numerical values	0 – 1	R1A	LT
	Acc_Reco	Recognition accuracy	Numerical, Percentage	0 – 100 %	R1A	LT

	Acc_Lat	Latency-aware F-score	Numerical, continuous	0 – 1	R1A	LT
Fundamental Frequency (F0) Analysis	Voicing Decision Error (VDE)	The proportion of frames for which an error of the voicing decision is made.	Numerical, Percentage	0 – 100 %	R1A	LT
	Gross Pitch Error (GPE)	The proportion of frames, where the decisions of both the pitch tracker and the ground truth are voiced, for which the relative error of F0 is higher than a threshold of 20%	Numerical, Percentage	0 – 100 %	R1A	LT
	F0 Frame Error (FFE)	The proportion of frames for which an error (either according to the GPE or the VDE criterion) is made.	Numerical, Percentage	0 – 100 %	R1A	LT
Respiration	Acc_ Resp	Inhalation detection F-score	Numerical, continuous	0 – 1	R1A	LT
	Acc_ Reco	Recognition accuracy for egressive and ingressive sounds	Numerical, Percentage	0 – 100 %	R1A	LT
Voice Tone	The dimensions of this have yet to be defined according to refined use-cases.					
Special Vocal Effect	The dimensions of this have yet to be defined according to refined use-cases.					
Doubling of Period	The dimensions of this have yet to be defined according to refined use-cases.					
Onset Detection Analysis	Acc_ Onset	Onset detection F-score: it is computed when considering as false the data falling outside a given threshold (typically of 25 ms)	Numerical, continuous	0 – 1	R1A	LT

Table 4-14. Assessment indices for Sound Processing module

4.2.6 Text-to-Song Module

A more elaborated speech synthesis system, the Text-to-Song Module will be one of the tools deployed in the Educational Processes. Its role will be to demonstrate rare singing style techniques using new lyrics, as well as replacing missing group members in the learning process.

The technical assessment of the module will be based on a series of indices described in the table below. The indices were chosen to show the system's ability to perform the tasks defined in the user requirements.

Assessment Indices

Module	**Text-to-Song**					
Assessment Category	**Assessment Indices**				**Requirements Examined**	**Assessment Phase**
	ID	**Description**	**Type**	**Values**		
Supported Inputs	MS	Ability to read a musical score with lyrics	Qualitative	Yes No	R24A	LT,CS, OA
	BMS	Ability to read a byzantine musical score	Qualitative	Yes No	R24A	LT,CS, OA
	MIDI	Ability to read a MIDI file with lyrics	Qualitative	Yes No	R24B	LT,CS, OA
	LNG	Number of supported languages	Numerical, Discrete	0..3	R24	LT,CS, OA
Musical Indices	VTH	Number of supported vocal techniques	Numerical, Discrete	0..Inf	R.24	LT,CS, OA
	TH	Number of available voices	Numerical, Discrete	0..10	R.24	LT,CS, OA
	STY	Number of supported singing styles	Numerical, Discrete	0..3	R.24	LT,CS, OA
Technical Indices	LT	System latency (time lapsed from reading the score until the first sample is output)	Numerical, Continuous	0..Inf (seconds)	Technical	LT

Table 4-15. Assessment indices for Text-to-Song module

4.3 Assessment of Data Fusion and Semantic Analysis

4.3.1 Ontology Engineering

4.3.1.1 Problem description and formulation

The ontology manifests the expert knowledge representation in the i-Treasures system. Thus, it has to be evaluated with respect to the degree of the expressivity and representation accuracy and the ability to reason under this knowledge. In order

to describe the assessment procedure, we must first define the meaning and the purpose of the ontology assessment task. Ontology assessment is the evaluation of its knowledge representation accuracy and the validation that the ontology represents this knowledge in an efficient, coherent and consistent manner. This is a crucial task, necessary before validating the ontology. Next, we describe the procedure that will be followed for the ontology assessment and the indices (measurers/metrics) that we plan to use for this purpose.

We use the following four general axes, proposed in [3], and their parameters for assessing the ontology:

1. Syntactic quality: a) correct syntax, b) syntactic richness
2. Semantic quality: a) consistency, b) ICT terms used
3. Pragmatic quality: a) comprehensiveness (class and relationship richness), b) accuracy, and c) relevance
4. Social quality: a) authority, b) history.

For our case, the fourth axis (social quality) is not of our concern, since it is not expected our ontology to be based on previous ontologies, given that the defined concepts are very specific. Regarding the syntactic quality axis, we focus on the correct syntax, which is the proper usage of the knowledge representation language and the syntactic richness, which refers to the number of syntactic features (e.g. classes, relationships, etc.). Regarding semantic quality, the logical consistency of the rules defined in the ontology is a crucial aspect of the quality, while the terms used to defined classes and relationships should be taken by a standard term vocabulary, e.g., the OSRD document. Thus, assessing in terms of the syntactic and semantic quality, the ontology shall be evaluated with respect to the correct, for the syntactic case, and the efficient, for the semantic case, utilization of description logic knowledge expression arsenal.

The pragmatic quality, contrary to the syntactic and semantic quality, represents the usefulness of the ontology to the end user. Thus, for assuring a high pragmatic quality, the ontology shall be evaluated with respect to the class and relationship number, the degree of the expert knowledge representation accuracy and relevance of the manifested knowledge in the ontology with the domain knowledge.

4.3.1.2 Ontology assessment criteria and indices

Assessing the ontology in terms of the syntactic quality entails the detection of syntactic errors, in order to be corrected, potentially occurred during the development of the ontology, using the employed knowledge representation language (description logic (DL) in our case). However, we do not plan to follow a specific assessment procedure for validating the conformance of the ontology to this requirement. This is because the software used for the ontology development contains reasoners that perform automatically this validation task. However, we can ensure a syntactic richness by demanding the ontology to utilize as much DL expressivity as possible (DL extensions, such as functional properties, inverse properties, role hierarchies, etc.). Thus, the number of the utilized syntactical rules is a proper index for this end.

The consistency aspect of the semantic quality can be assessed by counting the entities defined in the ontology with ambiguous names, or in other words, the number of entities with a name similar to the name of another semantically irrelevant entity. Thus, the number of inconsistencies can be used as the assessment index for this purpose. To ensure a high semantic quality with respect to consistency, we aim to build ontology with minimal, if not zero, semantic inconsistencies. Moreover, high semantic quality needs, in our case, the usage of terms defined in the ontology

terminology, and thus, this number is also a useful index for semantic quality assessment.

For the pragmatic quality assessment, metrics related to the ontology richness with respect to the depth and breadth of the hierarchical class tree can be used. First, let us define the required quantities for defining the measures [4], [5]:

- *Depth* is the number of nodes of a path in a tree
- *Breadth* is the number of nodes of a specific level in a tree,
- *Branching factor* of a node in a tree is the number of its children,

where with the term “tree” we refer to the hierarchical (tree) structure of the class defined in the ontology. Also, in what follows, class tree is the sub-tree of the whole ontology tree, where the root node is a specific class. The following indices can be used for the assessment of the ontology with respect to class richness.

1. Class tree depth: a) maximum depth, b) minimum depth, c) average depth.
2. Class tree breadth: a) maximum breadth, b) minimum breadth, c) average breadth
3. Tree branching factor: a) maximum branching factor, b) minimum branching factor, and c) average branching factor.

Moreover, for the pragmatic quality assessment, the following measures provide us with the ability to evaluate the ontology with respect to the class and relationship richness [4], [5]:

4. Schema metrics have to do with the ontological schema efficiency assessment. The indices are: a) Attribute richness : #attributes in all classes / #classes, b) Relationships richness: #relations / (#subclasses+#relations).
5. Instance metrics has to do with the assessment of the ontology knowledge base: a) Class richness: #class used/#class defined, b) Average population: #instances/#classes, and c) Cohesion: #separate relationship graph components
6. Class metrics are metrics defined for each class separately: a) Importance:#instances of class and subclasses / #total instances, b) Inheritance richness of a class C: average of subclasses per class that are descendants of C, c) Relationship richness of a class C: #instances of relations (properties) of C with another class / #relations including C, and d) Connectivity of a class C: #instances of other classes connected to instances of that class via any relationship.

Regarding the assessment of the ontology knowledge representation accuracy and relevance to the expert knowledge, (i.e., the accuracy and relevance parameters of the pragmatic quality), there are no absolute measures in terms of which this assessment can take place. This type of assessment, which is qualitative in nature, should be undertaken by the domain knowledge experts and/or the ontology development team. The main output of this task will be a report about the degree of the knowledge representation accuracy and the proportion of expert knowledge eventually managed to be expressed with the ontology. The main differentiation of this type of assessment from others is that it does not contain measurable quantities.

Ontology Engineering						
Assessment Category	**Assessment Indices**				**Requirements Examined**	**Assessment Phase**
	ID	**Description**	**Type**	**Values**		
Ontology assessment	PA1	Class tree depth (maximum, minimum average)	Numerical-Discrete	0-NC	R12A - E	LT
	PA2	Class tree breadth (maximum minimum average)	Numerical-Discrete	0-NC	R12A - E	LT
	PA3	Tree branching factor	Numerical-Discrete	0-NC	R12A - E	LT
	PA4	Attribute richness: #attributes in all classes / #classes.	Numerical-continuous	0-1	R12A - E	LT
	PA5	Relationships richness: #relations / (#subclasses+#relations)				
	PA6	Class richness: #class used/#class defined	Numerical-continuous	0-1	R12A - E	LT
	PA7	Average population: #instances/#classes	Numerical-continuous	0-1		
	PA8	Cohesion: #separate relationship graph components	Numerical-discrete	0-…		
	PA9	Importance:#instances of class and subclasses / #total instances	Numerical-continuous	0-1	R12A - E	LT
	PA10	Inheritance richness of a class C: average of subclasses per class that are descendants of C.	Numerical-discrete	0-…	R12A - E	LT
	PA11	Relationship richness of a class C: #instances of relations (properties) of C with another class / #relations including C.	Numerical-continuous	0-1	R12A - E	LT
	PA12	Connectivity of a class C: #instances of other classes connected to instances of that class via any relationship.	Numerical-discrete	0-…	R12A - E	LT

Table 4-16. Assessment indices for Ontology Engineering

4.3.2 Multimodal Data Fusion Assessment Protocol

In this section, we provide the assessment protocol and the indices/metrics that shall be used for the assessment of the data fusion. Fusion is a process that is a part of the probabilistic modeling and inference of ICH content, where high level concepts are detected and identified based on (medium-level) features extracted in a previous phase. Since in our case we deal with multiple modalities, we inevitably have to deal with multimodal fusion. Of course, it is crucial to select the proper features, which are extracted from various modalities, to be fused together. Thus, an appropriate assessment methodology and the corresponding indices for the assessment of the fusion process must be defined.

In the context of i-Treasures our intention is to rely on probabilistic models (e.g. multi-entity Bayesian network, MEBN) in order to fuse multimodal information and infer the unknown ICH concepts. Specifically, we aim to perform classification using features from multiple modalities and thus, perform mapping of the features to a common probability space, i.e., a space where they can be meaningfully fused with each other. In practice, this is achieved by modeling the feature values as random variables that participate jointly in a distribution (defined, e.g., by the MEBN). In assessing the efficiency of the fusion task, we need to evaluate the impact of using a feature or a set of features in combinations with others. We can perform the assessment by testing whether the dependencies between features are realistic and justified by the test data set. Before we proceed to the definition of the indices that can be used for the data fusion assessment procedure, we describe in the following list the assessment cases in which the indices will be employed:

1) two features belonging to different modalities,
2) two sets of features, where each set contains features of a single modality, different than that of the other set,
3) two sets of features, where each set contains features of a group of modalities.

Moreover, we need to state that a ground-truth dataset is necessary for evaluating the below mentioned indices.

As data fusion assessment indices, we plan to employ the mutual information measure [6] and other variants, in order to evaluate the degree of dependency (or independency) between the random variables of our probabilistic model and validate our dependency assumptions. Mutual information (MI), $I(X;Y)$ indicates the degree of dependency between two variables, X,Y. It is in essence the *Kullback-Leibler* divergence, D_{KL}, (or else, the *relative entropy*) between the joint probability distribution of X and Y, $p(X,Y)$ and the product of their distributions $q(X,Y) = p(X)\,p(Y)$:

$$I(X;Y) = D_{KL}(p \,\|\, q) = \sum_{X}\sum_{Y} p(X,Y)\log\frac{(X,Y)}{q(X,Y)} \qquad (4.1).$$

If X and Y are independent, MI becomes zero.

For the assessment procedure, X and Y shall be medium level features belonging to the same or different modality. The purpose here is to assess whether the inclusion of a specific feature contributes in further improving the expected performance of the semantic analysis task.

Another index that we plan to use, based also on the mutual information concept, is the distance metric of mutual information, defined as

$$D(X,Y) = H(X,Y) - I(Y;X) \quad (4.2),$$

where $H(X,Y)$ is the entropy of $p(X,Y)$. The purpose of this quantity is to provide a proper metric based on the mutual information that satisfies the necessary and sufficient conditions for a measure to be characterized as a metric.

Finally, we also plan to use the *total correlation (TC)*, which is a multivariate generalization of mutual information. It measures the dependency (or independency) between two set of variables $X_S = \{X_1, X_2, \ldots, X_n\}$ and $Y_S = \{Y_1, Y_2, \ldots, Y_m\}$, where *m,n* are positive integers. *TC* is (similarly with *MI*) the *KL* divergence between $p(X_S, Y_S)$ and $q(X_S, Y_S) = p(X_S)\,p(Y_S)$

$$TC(X_s;Y_s) = D_{KL}(p \,\|\, q) = \sum_{X_1} \cdots \sum_{Y_1} \cdots \sum_{Y_s} p(X_s, Y_s) \log \frac{(X_s, Y_s)}{q(X_s, Y_s)} \quad (4.3).$$

An important issue is that for the assessment of the indices we need explicitly the probability distributions p and q. The most straightforward way to do this is to assume that the probabilities have an known analytical form convenient for computations (e.g., Gaussian for the continuous case and multinomial for the discrete case), estimate the necessary parameters using a ground truth dataset, and evaluate the indices using their analytical form.

Fortunately, however, our probabilistic formulation of the semantic analysis and data fusion problem (i.e., the use of (multi-entity) Bayesian Networks to model the distributions) allow us to use the marginal distribution of the random variables participating in the joint distribution of the Bayesian network (defined as the p and q distributions). This of course makes necessary the training of the Bayesian network, i.e., the estimation of the joint distribution of all model variables, something that is not necessary for the simple cases mentioned previously.

Multimodal Data Fusion						
Assessment Category	**Assessment Indices**				**Requirements Examined**	**Assessment Phase**
	ID	**Description**	**Type**	**Values**		
Multimodal Data Fusion	MF1	Mutual information, I(X;Y) indicates the degree of (in)dependency between two variables, X ,Y. It becomes zero when the random variables are independent.	Numerical, continuous	0-inf.	R12A - E	LT
	MF2	Distance metric of mutual information, defined as D(X,Y)=H(X,Y)+I(Y;X). The purpose of this index is to provide a mutual information based metric.	Numerical, continuous	0-inf.	R12A - E	LT
	MF3	Total correlation is the multivariate	Numerical,	0-inf.	R12A - E	LT

		generalization of mutual information. It is measures the dependency (or independency) between two set of variables XS={X1,X2,X3…} and YS={Y1,Y2,Y3,…}. It is in essence the *KL* divergence between p(XS,YS) and p(XS)p(YS).	continuous			

Table 4-17. Assessment indices for Multimodal Data Fusion

4.3.3 Semantic Analysis and Classification

4.3.3.1 Problem formulation

The main goal of the semantic analysis module is to classify various entities, e.g., dance and singing styles, found in ICH content, to predefined classes. These classes are concepts of the domain knowledge defined rigorously in the ontology. Note that the classification shall be based on information from multiple modalities. Thus, annotation corresponds to assigning known concept labels to unclassified entities.

The classification problem is formally defined as assigning automatically a class label to object (entities found in the ICH content). The classes and their number are known a priori. Also the classifier, i.e., the method that assigns labels to entities, is automated. In this way, the unknown entities are classified. In the assessment framework, the experimental assessment protocol that should be followed for the assessment of the semantic analysis algorithms is defined in the next section. The classifiers must be assessed in terms of indices (metrics) that typically rely on the number of correctly (or not) classified cases among the total number of cases. Thus, the performance assessment of the classification needs a ground truth test data set, where the class labels are known (the entities have been annotated by human experts). Moreover, metrics are needed to evaluate quantitatively the efficiency of the classifier. Next, we describe the assessment protocol and the corresponding performance assessment metrics.

4.3.3.2 Classification Assessment Protocol

At the core of the semantic analysis task rests the modeling of the high and medium level concepts (features) and the relationships with each other. This type of modeling will be probabilistic. This means that probability distributions will be used to model the correlation and dependency between a high and one or more medium level features, as well as dependencies between high level features. Our intention is to use Multi-entity Bayesian Networks [7] for these modeling purposes, which is a probabilistic model extended to incorporate also first-order logic. The main feature of this type of modeling is the ability to create situation specific Bayesian Networks, depending on the situation that arises. For example, the number of steps, dancers, etc., may vary per performance. Thus, each dance related entity (e.g., step, dancer) should be modeled by a different variable and the number of variables may vary per situation.

The outcome of semantic analysis will be the labeling of the examined entities, which essentially a classification problem. This in practice means that random variables will

model the class labels and their inferred distribution based on the observations will help to classify the objects by assigning the most probable label to them.

The efficiency of the classification task is expressed qualitatively by measures that take into account the correctly and/or incorrectly classification results. In classification, a class is assigned to an entity. Let *N* be the number of all entities. Then, in order to use the following metrics, we need to use a test data set (with ground truth annotations) and define for each class the following quantities:

- number of *true-positives (tp),* i.e., the number of the entities of the test data set classified correctly to the class,
- number of *true-negatives (tn),* i.e., the number of entities that correctly have not been classified to the class,
- number of false positives (*fp),* i.e., the number of entities that incorrectly have been classified to the class and
- number of false negatives (*fn)* , i.e., the number of entities that incorrectly have not been classified to the class

Based on these quantities, the following metrics can be used [8]: a) ***Precision*** (or positive predictive value) is high when less false positives appear in the classification, b) ***Recall*** (or true positive rate), which is high when less false negatives appear in the classification, c) ***F-measure***, which is the harmonic mean of precision and recall and is typically used to quantify the performance of a system using a single index, and d) ***Average precision***, which is a measure that is based on the behavior of the precision measure as a function of the recall.

Depending on the nature of the problem at hand, we will employ the most appropriate subset of the aforementioned indices. Finally, for assessing the classification performance as “satisfactory”, all measures should be above the baseline approach for all conducted experiments, using a test data set with *ground truth* annotations. In this way, the calculation of the true/false positives/negatives will be feasible, since the class of each entity we aim to classify is a priori known.

Semantic Analysis						
Assessment Category	**Assessment Indices**				**Requirements Examined**	**Assessment Phase**
	ID	**Description**	**Type**	**Values**		
Classification	PREC	*Precision=#true-positive/(#true-positive+#false-positive)* Precision is high when less *false positives* appear in the classification and low in the opposite case.	Numerical-continuous	0-1	-	LT
	REC	*Recall=#true-positive/(#true-positive+#false-negative)*	Numerical-Continuous	0-1	-	LT

		Recall is high when less negative positives appear in the classification, while a lower precision corresponds to more negative positives				
	F	*F-measure* is the harmonic mean of precision and recall. Thus, it combines both of the previous indices	Numerical-Continuous	0-inf	-	LT
	AP	*Average precision* is the average of every possible value of the precision metric which explicitly depends on the recall. At each precision value, recall also varies	Numerical-Continuous	0-1	-	LT

Table 4-18. Assessment indices for Semantic Analysis

4.4 Educational Process and Platform Interface Assessment

4.4.1 Educational Processes

The educational role of i-Treasures project is based on the design and development of versatile, adaptive and well-structured learning processes and educational scenarios according to the needs of the users. The implementation of this core functionality of i-Treasures platform will be based on open source Learning Management System (LMS) solutions that will provide the appropriate tools to realize the educational scenarios that will be designed.

The assessment of educational processes from a technical perspective will be based on various technical indices (the table below contains some of the main indices considered). These indices will be handled in order to form an overall conformity index that will convey the compliance rate of the developed educational platform to the general LMS model principles. Moreover, each learning activity (as this is designed through the i-Treasures Pedagogical Planner – see D5.1 for more details), will be evaluated by considering data about students' participation (e.g. number of accesses to resources, level of engagement during the activity, etc.).

The evaluation of the educational processes from the users' point of view will be performed within the case studies, described in Section 5.

Educational platform						
Assessment Category	**Assessment Indices**				**Requirements Examined**	**Assessment Phase**
	ID	**Description**	**Type**	**Values**		
Multi-method learning	MML 1	Availability of learning materials in various formats (audio, video, text, images, etc.)	Qualitative	Audio Video Images Text …	R17A, R17G, R22A, R22B	LT
	MML 2	Availability of testing tools (e.g. quizzes)	Qualitative	Yes No	R17C R17E	LT
	MML 3	Availability of Text-to-Song tool	Qualitative	Yes No	-	LT
	MML 4	Availability of video annotation tool	Qualitative	Yes No	-	LT
	MML 5	Availability of annotated musical score	Qualitative	Yes No	-	LT
Learning process adjustment	LPA1	Possibility to set/ choose topics	Qualitative	Yes No	R18B	LT
	LPA2	Possibility to set /choose difficulty levels	Qualitative	Beginners Intermediate Advanced	R25D	LT
	LPA3	Possibility to set /choose contexts	Qualitative	Formal Informal ….	-	LT
	LPA4	Possibility to offer recovery activities according to responses to quizzes	Qualitative	Yes No	R18A	LT
	LP5	Availability of tracking functionalities	Qualitative	Yes No	--	LT
Users' interactions	UI1	Possibility to interact with others	Qualitative	Email Forum Chat …	R23A, R23B, R43A	LT
	UI2	Possibility to work/learn in group	Qualitative	Wiki Googledocs ……	R23A,	LT
Performance	PD1	Availability of musical instruments	Qualitative	Yes No	R27A	CS,OA

decomposition		separation				
	PD2	Availability of voices separation	Qualitative	Yes No	R27B	CS,OA
Text to song functionality	TS1	Availability of entering marks and lyrics	Qualitative	Yes No	R24A	CS,OA
	TS2	Availability of entering modern musical annotation	Qualitative	Yes No	R24B	CS,OA

Table 4-19. Assessment indices for Educational Platform

4.4.2 3D Visualisation Module

The i-Treasures platform will use 3D visualization techniques (3D avatars) and tools to support people in learning or mastering the different types of ICH. In order to achieve this goal, a web-based game like application will be developed. To this end, the application will contain a virtual tutor who gives instructions and evaluates user inputs. The proposed game-like application is expected to help those who want to learn the basics or to make practice.

In this vein, the assessment of the 3D visualization module will focus mainly on three dimensions, i.e., avatar specifications, user's performance feedback and personalization of learning activity. The table below presents the major assessment indices that will reflect the quality of the 3D module for sensorimotor learning.

3D Visualisation Module						
Assessment Category	**Assessment Indices**				**Requirements Examined**	**Assessment Phase**
	ID	**Description**	**Type**	**Values**		
Avatar Specifications	RotVert	The range of rotation angles of the avatar around the vertical axis.	Quantitative, continuous	-180^{o} - 180^{o}	R28A R28G R33A	LT
	RotHor	The range of rotation angles of the avatar around the horizontal axis.	Quantitative, continuous	-180^{o} - 180^{o}	R28A R28G R33A	LT
	ZoomF	The zoom factor available for the avatar	Quantitative, Discrete	0.1 - Inf (0.1 step)	R33A	LT
	LagTime	Time interval between user's action and avatar's response	Quantitative, Continuous	0 - Inf	-	LT
Feedback	Fb1	Availability of audio/visual feedback	Qualitative	Yes No	R31A	LT
	Fb2	Availability of posture and gesture feedback	Qualitative	Yes No	R31B R31D	LT

	Fb3	Availability of vocal tract feedback	Qualitative	Yes No	R31C	LT
	Fb4	Availability of student's performance annotation and reproduction	Qualitative	Yes No	R32A R34B R35A	LT
Personalization	Pers 1	Difficulty level settings for sensorimotor learning	Qualitative	Manual Dynamic None	R30C R30D R30E	LT
	Pers 2	Granular reproduction of learning content (e.g. specific movements, gestures, postures or overall complex artistic performance)	Qualitative	None Low Medium High	R30A R30B R30D	LT
Intangible Musical Instrument	IMI	Granularity of audio excerpts database (e.g. style, rhythm, pitch, type)	Qualitative	Low Medium High	R36A R36B R36C	LT

Table 4-20. Assessment indices for Educational Platform

4.4.3 Web Platform Interface

In the context of i-Treasures project, the objective of the Web Platform Interface Assessment is based on the following steps:

1. Questionnaire: Development of a structured questionnaire based on the criteria reported to the ISO/IEC 25010. Specifically, the questionnaire will make use of major criteria and sub-criteria. The latter constitute the various dimensions that reflect the major criteria. Each of the sub-criteria will be assessed by specific explanatory questions, expressed in 5-level scales from '1' = completely unacceptable/unclear/irrelevant to '5' = completely acceptable/clear/relevant. Among the demographics of the questionnaire emphasis will be given between experts and ordinary users.

2. Pilot: In this step a small number of experts and users will be asked to evaluate the content, the meaning and the wording of the questionnaire. The original questionnaire will be corrected / amended according to the reactions of the pilot experts / users.

3. Survey: The final questionnaire will be distributed to a larger number of experts and representative users. Emphasis will be given to the protocol used for the sampling methodology.

4. **Consistency:** Consistency of the survey instrument will be assessed with various statistical methods such as construct internal consistency, construct validity, construct composite reliability, and construct discriminant reliability. Additionally, tests investigating "common method bias" will be employed.

5. **Analysis:** The analysis is based on two phases. The first, detects the weights that should be assigned to each major criterion, sub-criterion, and specific question. These weights may be achieved by applying exploratory factor analysis (EFA), and will be used for calculating the indices that reflect the assessment of the Web

Platform Interface. Thus, the scope of this phase is to develop indices which reflect the latent variables of each group of criteria (first order and second order). Although the values of the questions are discrete between 1 and 5, the values of the latent variables are continuous between 1 and 5. The scope of the second phase is to investigate the various relationships between the latent variables developed in phase one. This phase is important because it indicates the cause and effect relationships between the various qualities of the Web Platform Interface.

6. Changing: The investigation of the latent variables explaining the criteria and sub-criteria and the relationships among them will be used to correct/amend various entities of the Web Platform Interface.

Web Platform Interface						
Assessment Category	**Assessment Indices**				**Requirements Examined**	**Assessment Phase**
	ID	**Description**	**Type**	**Values**		
Goals of the platform	G1	Clarity of the platform's goals	Numerical -Discrete	1-5		LT
	G2	Clarity of the institutions/organizations responsible for the platform				LT
Content	CON1	Content quality/ /representativeness /usefulness of different ICH forms			R39	LT
	CON2	Content accuracy/validity				LT
	CON3	Content understandability/ organization				LT
	CON4	Is the content outdate/obsolete?				LT
	CON5	Content uniqueness LT				LT
	CON6	Multimedia content quality				LT
Browsing ability	BA1	Availability of search mechanisms?			R41	LT
	BA2	Site map availability				LT
	BA3	Platform locationl awareness of the user (does user to know in which part of the site he/she is at all times)?				LT

	BA 4	Search engine discoverability				LT
	BA 5	Hyperlink identification and clarity	Numerical -Discrete	1-5		LT
	BA 6	Platform appearance consistency across sections				LT
Appearance	AP P1	Image quality/resolution				LT
	AP P2	Color combination balance				LT
	AP P3	Font quality				LT
Interaction	INT 1	QA section availability	Boolean	Yes/No		LT
	INT 2	Contact form availability				LT
Customization	CU 1	Interface adaptivity based on user role			R38A	LT
	CU 2	Font size changeability on demand				LT

Table 4-21. Assessment indices for Web Platform

Additionally we will use the indices based on the non-functional requirements that are presented in Section 4.1 and categorized according to Tables 4-1 to 4-9.

5. Use Cases Evaluation

5.1 Usability Evaluation through Case Studies

Use cases evaluation phase refers to the evaluation of the developed platform and its independent components in real conditions. The main objective of this assessment phase is the evaluation of platform's usability. Usability is a crucial characteristic of i-Treasures project since its target is to preserve and transmit knowledge to a wide range of users.

Through time many definitions for usability have been proposed. Two of the most established definitions can be found in international standard for the evaluation of software ISO 9241-11 [9] and ISO 9126 [10]. ISO 9241-11 defines usability as "*the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use*". In ISO 9126, usability is defined as "*the capability of the software product to be understood, learned, used and attractive to the user, when used under specified conditions*". In other words, usability studies relate to evaluating a product by testing it on representative users while they focus not only on how well users can learn and use a product to achieve their goals but also on how satisfied users are with that process. This can be seen as an irreplaceable usability practice since it gives direct input on how real users use the system. Usability studies examine three principles: effectiveness, efficiency and overall satisfaction of the user.

- Effectiveness is the capability of the product to enable users to achieve specified goals with accuracy and completeness in a specified context of use.
- Efficiency is the capability of the product to enable users to expend appropriate amounts of resources in relation to the effectiveness achieved in a specified context of use.
- Satisfaction is the capability of the product to adequately satisfy users in a specified context of use.

In this context, usability evaluation will be performed through a series of use case evaluation tests that will be carried out by organising case studies. The case studies will engage an adequate number of real users, both experts and learners so as to extract valid conclusions. The objectives of the use cases evaluation are mainly based on the requirements of the potential users of the platform that are identified within deliverable D2.1.

5.2 Usability Data Collection Methods

It is common during a usability study participants to try to complete typical tasks while observers watch, listen and take notes. The goal is to identify any usability problems, collect qualitative (that approximate or characterize but does not measure the attributes, properties, and characteristics of a thing or a phenomenon) and quantitative (that quantify and verify the attributes, properties, and characteristics of a thing or a phenomenon) data and better understand the users' satisfaction with the product and their motivations/perceptions in addition to their actions. The methods popularly used to gather usability data can be divided into two categories, namely testing and inquiry, and are described below.

5.2.1 Testing

In usability testing approach, representative users work on typical tasks using the system (or the prototype) and the evaluators use the results to understand how the user interface as well as the system in general supports the users to perform their tasks. The most popular techniques used to gather data during a usability test are the following.

5.2.1.1 Think Aloud Protocol

Think Aloud Protocol was introduced in the usability field by Clayton Lewis [11]**Error! Reference source not found.** and was based on the techniques of protocol analysis by Ericsson and Simon [12]. Think Aloud Protocol involves participants thinking aloud as they are performing a set of specified tasks. During the course of a usability test, the test users are asked to verbalize their movements, thoughts, feelings, and opinions while interacting with the system. That is the reason why it is also referred as Concurrent Think Aloud Protocol so as to differentiate it from Retrospective Think Aloud described in Section 5.2.1.2. More specifically, the test users are provided with the product to be tested and a set of tasks to perform. Then, they are asked to perform the tasks using the product and explain what they are thinking about while working with the product's interface. Observers of such a test are asked to objectively take notes of everything that users say, without attempting to interpret their actions and words. Test sessions are often audio- and video-recorded so that developers can go back and refer to what participants did and how they reacted. The purpose of this method is to make explicit what is implicitly present in subjects who are able to perform a specific task.

Thinking aloud is very useful in capturing a wide range of cognitive activities and enables observers to see first-hand the process of task completion (rather than only its final product). Furthermore, allows observers to understand how the user approaches the interface and what considerations the user keeps in mind when using the product. If the user expresses that the sequence of steps dictated by the product to accomplish their task goal is different from what they expected, perhaps the interface is convoluted. Although the main benefit of the thinking aloud protocol is a better understanding of the user's mental model and interaction with the product, there are other benefits as well. For example, the terminology the user uses to express an idea or function should be incorporated into the product design or at least into its documentation. However, the main drawbacks of Thinking Aloud Protocol are the non-natural environment of the testing process to the user and the inability to capture quantitative data.

Usability principles covered:

Effectiveness: Yes, Efficiency: No, Satisfaction: Yes

5.2.1.2 Retrospective Testing

Retrospective Testing or Retrospective Think Aloud is a form of Think Aloud Protocol that is performed after the user testing session activities instead of during them. Fairly often the retrospective protocol is stimulated by using a visual reminder such as a video replay. If a video replay of the usability test session is available, the observers can collect more information by reviewing the replay together with the user participants and asking them questions regarding their behaviour during the test. Consequently, this technique should be used along with other techniques, especially those where the interaction between the observers and the participants is restricted. Moreover, both quantitative and qualitative data can be collected while in concurrent thinking aloud quantitative information gathering is not an option. However, in

retrospective testing each test session lasts at least twice as long. Another obvious requirement for using this technique is that the user's interaction with the computer needs to be recorded and replayed.

Usability principles covered:

Effectiveness: Yes, Efficiency: Yes, Satisfaction: Yes

5.2.1.3 Co-discovery Learning

Co-discovery Learning is an adaptation of the most commonly used Think Aloud Protocol. In Co-discovery Learning, users are grouped in pairs and attempt to perform tasks together by talking aloud naturally to each other whilst being observed. They are to help each other in the same manner as they would if they were working together towards accomplishing a common goal using the product. They are encouraged to explain what they are thinking about while working on the tasks. Compared to Think Aloud Protocol, this technique makes it more natural for the test users to verbalize their thoughts during the test while retaining the great facilities of thinking aloud, pursuing of the users train of thought and notating erroneous assumptions about the system. It is also optimal to pair users who know each other so that they do not feel uncomfortable working together.

Co-discovery Learning is more realistic than a single user scenario, as people in work environments often work in teams. The users often find it easier and more natural to vocalize thoughts with a colleague present. The evaluators can also quantify the time taken for various tasks, the number of tasks competed correctly, the error frequency, numbers of times the users accessed the help system etc. [13]. These observations can form the ground to make more qualitative judgments such as the success or lack of the entire system, system sub-components, help system, effort required to achieve a particular result and quality of interface.

Usability principles covered:

Effectiveness: Yes, Efficiency: No, Satisfaction: Yes

5.2.1.4 Eye-tracking

Eye tracking is the process of measuring either the point of gaze (where one is looking) or the motion of an eye relative to the head. Eye movement is typically divided into fixations and saccades – when the eye gaze pauses in a certain position, and when it moves to another position, respectively. The resulting series of fixations and saccades is called a scanpath. Scanpaths are useful for analyzing cognitive intent, interest, and salience while eye tracking in human-computer interaction (HCI) typically investigates the scanpaths for usability purposes.

There are numerous eye tracking techniques but the most popular and widely used are video-based eye trackers. A camera focuses on one or both eyes and records their movement as the viewer looks at some kind of stimulus. Most modern eye-trackers use the centre of the pupil and infrared / near-infrared non-collimated light to create corneal reflections. The vector between the pupil centre and the corneal reflections can be used to compute the point of regard on surface or the gaze direction. A simple calibration procedure of the individual is usually needed before using the eye tracker.

A wide variety of disciplines use eye tracking techniques, including cognitive science, psychology, HCI, marketing research, and medical research. Specific applications include the tracking eye movement in language reading, music reading, human activity recognition, the perception of advertising, and the playing of sports. More

recently, eye tracking has become a key method to test usability of software. While traditional usability techniques are often quite powerful in providing information on clicking and scrolling patterns, eye tracking augments traditional usability methods by providing additional indisputable, objective and convincing data describing behaviour and usability problems that the test participant cannot report and the researcher cannot observe. More specifically, it provides observers and testers with the ability to analyze user interaction between the clicks, how much time a user spends between clicks and unique information about first glance, search patterns and failed search. Eye tracking can be used together with a variety of research methods, including observations, interviews and Think Aloud Protocols. As a result it may yield valuable insight into which features are the most eye-catching, cause confusion or be ignored altogether as well as facilitate the assessment of search efficiency, navigation usability, distinctiveness, attractiveness and overall design.

Usability principles covered:

Effectiveness: No, Efficiency: Yes, Satisfaction: No

5.2.2 Inquiry

During usability test, evaluators need to obtain information about users' likes, dislikes, needs and understanding of the system by talking to them, observing them or letting them answer questions verbally or in written form. The inquiry data collection methods can be divided into two categories, the traditional ones and the modern software-based ones.

5.2.2.1 Traditional approach

Some of the most common and useful traditional usability inquiry methods are the following.

Field Observation

Field Observation involves the visit of the usability evaluators to the users' workplace and observation of their work in order to understand how the users use the system to accomplish their tasks, if they use the system the way expected and what kind of mental model the users have about the system. However, field observation is time consuming, there is usually insufficient number of observations and the presence of observers may alter the behaviour of the users and the working procedure in general.

Usability principles covered:

Effectiveness: Yes, Efficiency: No, Satisfaction: Yes

Focus Groups

This is a data collecting technique where about 6 to 9 users are brought together to discuss issues relating to the system. A usability evaluator plays the role of a moderator, who needs to prepare the list of issues to be discussed beforehand and seek to gather the needed information from the discussion. This can capture spontaneous user reactions and ideas that evolve in the dynamic group process. A serious consideration about Focus Groups technique is the skilfulness of the moderator who needs to be experienced in group facilitation and communication to make a focus group successful. It is not as simple as preparing questions since moderator needs to facilitate and guide discussion in real time. In addition, the data collected may possibly be biased, have low validity and be difficult to analyze

because of their unstructured free-flowing nature and participants' inability to be candid.

Usability principles covered:

Effectiveness: Yes, Efficiency: No, Satisfaction: Yes

Questionnaires

Questionnaires have long been used to evaluate products since they provide answers to a variety of questions according to the needs. Moreover, questionnaires can be answered anonymously, allow time before responding, can be administered to many users at distant sites simultaneously and impose uniformity by asking all respondents the same questions. On the other hand, people can often express themselves better orally than in writing and informative questions take time to be developed and are not as flexible as interviews.

Questionnaires can be either "home grown" or measure against a benchmark of the use of standardized and publicly available surveys such as SUMI and WAMMI which are marked against a database of previous usability measurements. SUMI (University College Cork) is a brief questionnaire that is marked against a benchmark of responses to surveys of systems. WAMMI is an on-line survey administered as a page on the web site and users are asked to complete it before they leave the page. This gives ongoing feedback to continue monitoring how the web site is used. Each organization using the SUMI or WAMMI surveys send back their results to the Human Factor Research Group (HFRG) who provides statistical results from the database build of all SUMI/WAMMI users. Other questionnaires specifically designed to access aspects of usability, the validity and/or reliability are the following: QUIS (Questionnaire for User Interface Satisfaction) developed by University of Maryland, PUEU (Perceived Usefulness and Ease of Use) developed by IBM, CSUQ (Computer System Usability Questionnaire) developed by IBM and PUTQ (Purdue Usability Testing Questionnaire) developed by Purdue University.

Usability principles covered:

Effectiveness: Yes, Efficiency: No, Satisfaction: Yes

Interviews

In this technique, usability observers formulate questions about the product based on the issues of interest. Then, they interview representative users to ask them these questions in order to gather the information desired. Interviews are flexible, suitable to get in-depth information for sensitive topics and allow the interviewer to pursue unanticipated lines of inquiry. On the contrary, interviews are time consuming and sometimes the interviewer can unduly influence the responses of the interviewee. The methods of interviewing include unstructured interviewing and structured interviewing. Unstructured interviewing methods are used during the earlier stages of usability testing. The objective of the investigator at this stage is to gather as much information as possible concerning the user's experience. The interviewer does not have a well-defined agenda and is not concerned with any specific aspects of the system. The primary objective is to obtain information on procedures adopted by users and on their expectations of the system. Structured interviewing has a specific, predetermined agenda with specific questions to guide and direct the interview. Structured interviewing is more of an interrogation than unstructured interviewing, which is closer to a conversation.

A useful technique to obtain further information after the original questions are answered is the use of probes. Probes are used to encourage the subjects to

continue speaking, or to guide their response in a particular direction so a maximum amount of useful information is collected. Types of probes include:

- *Addition probe* encourages more information or clarifies certain responses from the test users. Either verbally or nonverbally the message is, "*Go on, tell me more*" or "*Don't stop*".
- *Reflecting probe*, by using a nondirective technique, encourages the test user to give more detailed information. The interviewer can reformulate the question or synthesize the previous response as a proposition.
- *Directive probe* specifies the direction in which a continuation of the reply should follow without suggesting any particular content. A directive probe may take the form of "*Why is the (the case)?*"
- *Defining probe* requires the subject to explain the meaning of a particular term or concept.

Usability principles covered:

Effectiveness: Yes, Efficiency: No, Satisfaction: Yes

5.2.2.2 Software-based approach

In recent years an increasing number of software tools involved in the usability evaluation process have emerged. These tools aim to automatically collect statistics about the detailed use of the examined system, a process called logging. Logging is useful because it shows how users perform their actual work and enables effortless automatically collection of data from a large number of users working under different circumstances. Typically, a software product log will contain data about distance covered by mouse cursor, speed of cursor, use of keyboard, use of mouse button, total time of user activity, the frequency with which each user has used each feature in the product and the frequency with which various events of interest (such as error messages) have occurred. Moreover, same logging tools enable the capturing of screenshots and videos, storage of user activities in log files and creation, storage and implementation of macros. Such information can be used to optimize frequently used features and to identify the features that are rarely used or not used. Statistics showing the frequency of various error situations and the use of online help can be used to improve the usability of future releases of the system by redesigning the features causing the most errors and most access for online help. Some of the most popular logging software approaches are shown in the next table.

#	**Software Name**	**Freeware**
01	Mousotron Pro 5.0	YES
02	Mouse Off-road 2.15	YES
03	Mini-Input 2.0	NO
04	Mouse Odometer 4.0	YES
05	Mouse Meter 1.51	NO
06	My Mouse Meter 1.0.9	YES
07	Mouse Clocker 1.0	YES
08	Exact Mouse 2.0	NO
09	Usability Logger 2.3	YES

10	321 Soft Screen Video Recorder 1.05	NO
11	Screen VidShot 2.2.0.14	NO
12	ZD Soft Screen Recorder 2.6.4.0	NO
13	Screen Video Recorder 1.5	NO
14	Screen Tracker 2.0	NO
15	Advanced Key and Mouse Recorder 2.80	NO
16	Action Mouse Mover 1.0	NO
17	Adamant Key Mouse Pro 3.3	NO
18	Axife Mouse Recorder 5.0.1	NO
19	ECTI 1.73	NO
20	Mouse Tamer 2.0	NO
21	Smack 1.06	NO
22	Mouse Machine 1.1	YES
23	Jitbit Macro Recorder 3.82	NO
24	Mouse Master 2.1	NO
25	Macro Wizard 4.1	NO

Table 5-1. Logging software for usability data collection

Usability principles covered:

Effectiveness: Yes, Efficiency: Yes, Satisfaction: No

5.3 Case Studies

5.3.1 Framework of case studies, data analysis and plan validity

As previously mentioned, the use cases evaluation phase is based on case studies that will evaluate the usability of the i-Treasures platform and its independent components. One case study is going to be applied per sub-use case since each sub-use case exhibits varying and distinct characteristics. The only exception is Tsamiko and Calus Dances that can be examined under the same case study. The case studies design will take into consideration the Runeson and Host researchers checklist [14] but will be suitably adapted to the needs of i-Treasures platform case. All functionalities of i-Treasures platform, from data capturing down to educational processes and data search and retrieval, will be assessed through the case studies.

The framework of the case studies is the following:

- Objective: The objective of a case study is defined according to the use case and sub-use case that is going to be examined and includes the evaluation of usability as well as the assessment of some major technical criteria.
- Methodology: The methodology engaged to realise the case study depends on the use case and sub-use case examined. The case study will include capturing, educational, and information retrieval scenarios in order to examine all the functionalities of i-Treasures platform. As far as the educational processes are concerned, since there are not clear, well defined and objective technical assessment criteria, the evaluation of learning activities will be based on users' (learners') final learning outcomes (performance to quizzes, quality of artifact/performance, etc.) while both students and

teachers will be asked to evaluate the effectiveness of each educational scenario, including adequacy of the proposed tools and resources. In the case of ICH data capturing and retrieval, technical assessment criteria are well defined and can be easily measured while the usability of such processes will be evaluated by using the aforementioned usability data collection techniques.

- Data collection tools: Usability data along with some critical technical information will be acquired using some of the data collection techniques analysed above.
- Type of actors: Each case study will employ a sufficient number of users, both experts and learners, in order to acquire ample information and elicit valid and meaningful conclusions.
- Requirements examined: Case studies will be designed so as to examine and evaluate as many user requirements as possible. In this direction, conducting all the case studies entails the evaluation of the entire set of user requirements.

The usability data extracted from the case studies will be analyzed using qualitative and quantitative techniques and significant user satisfaction indices will be derived. These indices are going to be used for the overall assessment of the use cases and the integrated platform. Moreover, web platform metrics analysis will take place. The users' interactions with i-Treasures platform will be tracked and information such as session duration, page views, users' trending, users' click locations etc will be analyzed and assessed. All these quantitative and qualitative data will reveal to what extent the identified user requirements were met and whether or not they have to be updated. In Appendix 7.2 the template of the case study evaluation report is provided.

The main reason for using case studies for validation purposes is that the examination of the data is most often conducted within the context of its use, that is, within the situation in which the activity takes place. Moreover, variations in terms of intrinsic, instrumental and collective approaches to case studies allow for both quantitative and qualitative analyses of the data. Last but not least, the detailed qualitative accounts often produced in case studies not only help to explore or describe the data in real-life environment, but also help to explain the complexities of real life situations which may not be captured through experimental or survey research.

5.3.2 Case Studies per Use Case

This section presents a concise but rather comprehensive draft of the case studies that are going to be conducted for each use case.

5.3.2.1 Rare Traditional Songs

Case Study 1	**ID**	RTS1	**Title**	Byzantine Music
Objectives	1) Assessment of recognition accuracy of chanting gestures 2) Assessment of recognition accuracy of articulators position 3) Assessment of recognition accuracy of chanting performance 4) Assessment of accuracy of the identification of different chanting styles 5) Assessment of the metric and rules used to evaluate singers			

	performance 6) Evaluation of usability and user-friendliness of platform for sensorimotor learning 7) Evaluation of usability and user-friendliness of web platform for research 8) Evaluation of educational scenarios 9) Assessment of the Text-to-Song module quality
Methodology	The use case will consist of different steps. One is related to the capture of chanting gestures, articulator positions and audio signal (singing voice). These parameters will be annotated by experts, and the body and gesture recognition, vocal tract sensing and modelling and sound processing modules will be tested based on these reference annotations. In a second step, the learning application will be tested by presenting it to new naïve users. The goal of this use case is to design an application which could be displayed in music schools, churches, or any other cultural institutions in order to give a brief introduction to chanting by learning one or a few simple chanting techniques, representative of these songs.
Data Collection Tools	1) Using metrics defined for Body Recognition Module, Vocal Tract Sensing and Modelling Module and Sound Processing Module assessment (Objectives 1, 2, 3) 2) Using Questionnaires (Objectives 4, 5, 6, 7, 8, 9) 3) Use of data analytics collected by the final system during the Demonstration phase
Type of Actors (Number)	2-5 experts and 10-40 learners
Requirements Examined (D2.1)	R1A, R1B, R3B, R3(G-H), R8F, R10A, R11C, R12D, R17C, R17E, R17(G-H), R22(A-B), R23A, R24A, R25(C-D), R26A, R28C, R29A, R30(A-E), R31A, R31C, R33D, R34B, R35A, R37(A-B), R38A, R39A, R40A, R40D, R41A, R41(D-E), R42A, R44A

Table 5-2. Overview of case study "Byzantine Music"

Case Study 2	**ID**	RTS2	**Title**	Canto a Tenore
Objectives	1) Assessment of recognition accuracy of singer's gestures 2) Assessment of recognition accuracy of articulators' position 3) Assessment of recognition accuracy of singer's performance 4) Assessment of accuracy of the identification of different styles/ voices/ modas, etc. 5) Assessment of the metric and rules used to evaluate singer's performance 6) Evaluation of usability and user-friendliness of platform for sensorimotor learning 7) Evaluation of usability and user-friendliness of web platform for			

	research 8) Evaluation of educational scenarios 9) Assessment of the Text-to-Song module quality
Methodology	The use case will consist of different steps. One is related to the capture of chanting gestures, articulator positions and audio signal (singing voice). These parameters will be annotated by experts, and the vocal tract sensing and modelling, body and gesture recognition and sound processing modules will be tested based on these reference annotations. In a second step, the learning application will be tested by presenting it to new naïve users. The goal of this use case is to design an application which could be displayed in music schools, or any other cultural institutions in order to give a brief introduction to Canto a Tenore by learning one or a few simple singing techniques, representative of these songs.
Data Collection Tools	1) Using metrics defined for Body Recognition Module, Vocal Tract Sensing and Modelling Module and Sound Processing Module assessment 2) Using Questionnaires 3) Use of data analytics collected by the final system during the Demonstration phase
Type of Actors (Number)	1-3 experts and 5-10 learners
Requirements Examined (D2.1)	R1.A, R2.A, (R3.A), R3.B, R3.G, (R3.H-I), (R4.A-B), (R5.A), (R8.F), R10.A, R11.C, (R12.D), (R14.A), R17.C, R17.E, R17.F, (R18.A), R22.A, R23.A, R24.A, R25.B, R25.C, R25.D, (R26.A), (R27.B), (R28.C), (R29.A), (R30A-E), R31.A, (R31.C), (R33.A), (R34.B), (R35.A), R37.A, (R37.B), R38.A, R39.B, R40.A, R40.D, (R41.A), R41.D, (R41E).

Table 5-3. Overview of case study "Canto a Tenore"

Case Study 3	**ID**	RTS3	**Title**	Cantu in Paghjella
Objectives	1) Assessment of recognition accuracy of articulators' position (tongue, lips, jaw) 2) Assessment of accuracy of the identification of different styles/ voices/ modas, etc. 3) Assessment of the metric and rules used to evaluate singer's performance 4) Evaluation of usability and user-friendliness of platform for sensorimotor learning 5) Evaluation of usability and user-friendliness of web platform for research 6) Evaluation of educational scenarios			

Methodology	The use case will consist of different parts. One is related to the capturing of singing and vocal technique movements, singing gestures, articulator positions and audio signal (singing voice). These parameters will be annotated by experts, and the vocal tract sensing and modelling, body and gesture recognition and sound processing modules will be tested based on these reference annotations. This procedure involves: 1) sensor assessment - calibration or pre-assessment for each sensor of the Hyper Helmet (HH), - data quality check or post-recording session for each sensor - synchronisation check 2) quality assessment - automatic check (eg zeros, saturation, amplitude range…) - manual check of randomly selected subsets 3) metrics - for each sensor, define metrics to assess articulators' positions - for each sensor, define metrics to assess articulators' gestures for each sound type (vowels, consonants…) for each singing unit (straight tone, melisma…) In the second part, the learning application will be tested by presenting it to new naïve users. The goal of this use case is to design an application which could be displayed in music schools, or any other cultural institutions in order to give a brief introduction to Cantu in Paghjella by learning one or a few simple singing techniques, representative of these songs.
Data Collection Tools	1) Using metrics defined for data collection with the HH at the level of the sensors, and software for synchronisation and visualisation 2) System modules + questionnaires (or interviews)
Type of Actors (Number)	1-3 experts and 5-10 learners
Requirements Examined (D2.1)	R1.A, R2.A, (R3.A), R3.B, R3.G, (R3.H-I), (R4.A-B), (R5.A), (R8.F), R10.A, R11.C, (R12.D), (R14.A), R17.C, R17.E, R17.F, (R18.A), R22.A, R23.A, R24.A, R25.B, R25.C, R25.D, (R26.A), (R27.B), (R28.C), (R29.A), (R30A-E), R31.A, (R31.C), (R33.A), (R34.B), (R35.A), R37.A, (R37.B), R38.A, R39.B, R40.A, R40.D, (R41.A), R41.D, (R41E).

Table 5-4. Overview of case study "Cantu in Paghjella"

Case Study 4	ID	RTS2	Title	Human BeatBox
Objectives	1) Assessment of recognition accuracy of articulators' position (tongue, lips, jaw) 2) Assessment of accuracy of the identification of different styles/ voices/ vocal techniques, etc. 3) Assessment of the metric and rules used to evaluate singer's performance 4) Evaluation of usability and user-friendliness of platform for sensorimotor learning 5) Evaluation of usability and user-friendliness of web platform for research 6) Evaluation of educational scenarios			
Methodology	The use case will consist of different parts. One is related to the capturing of singing and vocal technique movements, singing gestures, articulator positions and audio signal (singing voice). These parameters will be annotated by experts, and the vocal tract sensing and modelling, body and gesture recognition and sound processing modules will be tested based on these reference annotations. This procedure involves: 1) sensor assessment - calibration or pre-assessment for each sensor of the Hyper Helmet (HH), - data quality check or post-recording session for each sensor - synchronisation check 2) quality assessment - automatic check (eg zeros, saturation, amplitude range…) - manual check of randomly selected subsets 3) metrics - for each sensor, define metrics to assess articulators' positions - for each sensor, define metrics to assess articulators' gestures for each sound type (vowels, consonants…) for each singing unit (straight tone, melisma…) In the second part, the learning application will be tested by presenting it to new naïve users. The goal of this use case is to design an application which could be displayed in music schools, or any other cultural institutions in order to give a brief introduction to Human BeatBox by learning one or a few simple singing techniques, representative of these songs			
Data Collection Tools	1) Using metrics defined for data collection with the HH at the level of the sensors, and software for synchronisation and visualisation 2) System modules + questionnaires (or interviews)			

Type of Actors (Number)	1-3 experts and 5-10 learners
Requirements Examined (D2.1)	R1.A, R2.A, (R3.A), R3.B, R3.G, (R3.H-I), (R4.A-B), (R5.A), (R8.F), R10.A, R11.C, (R12.D), (R14.A), R17.C, R17.E, R17.F, (R18.A), R22.A, R23.A, R24.A, R25.B, R25.C, R25.D, (R26.A), (R27.B), (R28.C), (R29.A), (R30A-E), R31.A, (R31.C), (R33.A), (R34.B), (R35.A), R37.A, (R37.B), R38.A, R39.B, R40.A, R40.D, (R41.A), R41.D, (R41E).

Table 5-5. Overview of case study "Human BeatBox"

5.3.2.2 Rare Dance Interactions

Case Study 1	**ID**	RDI1	**Title**	Traditional Tsamiko/Calus Dancing
Objectives	1) Assessment of recognition accuracy of dance figures 2) Assessment of accuracy of the identification of different dance styles 3) Assessment of the metric and rules used to evaluate dancers performance 4) Evaluation of usability and user-friendliness of platform for sensorimotor learning 5) Evaluation of usability and user-friendliness of web platform for research 6) Evaluation of educational scenarios			
Methodology	The use case will consist of different steps. One is related to the capture of dance motions. These motions will be annotated by the experts, and the dance figure recognition module will be tested based on these reference annotations. In a second step, the learning application will be tested by presenting it to new naïve users. The goal of this use case is to design an application which could be displayed in museums/exhibitions/dance festivals/any other space, in order to give a brief introduction to these dances by learning one or a few basic steps, representative of these dances.			
Data Collection Tools	1) Using metrics defined for Body Recognition Module assessment (Objectives 1, 2) 2) Using Questionnaires (Objectives 3, 4, 5, 6) 3) Use of data analytics collected by the final system during the Demonstration phase			
Type of Actors (Number)	2-5 experts and 10-40 learners			

Requirements Examined (D2.1)	(R3C, R3D, R3E, R7(A-E), R8A, R9A, R10A, R11A,R11C, R13A,R13B, R16A, R17A, R17E, R17G, R18(A,B),R19(A-B), R20A, R21A, R22(A,B), R23(A,B), R(28-36), R(37-44)

Table 5-6. Overview of case study "Traditional Tsamiko/Calus Dancing"

Case Study 2	**ID**	RDI2	**Title**	Traditional Walloon Dancing
Objectives	1) Assessment of recognition accuracy of dance figures 2) Assessment of accuracy of the identification of different dance styles 3) Assessment of the metric and rules used to evaluate dancers performance 4) Evaluation of usability and user-friendliness of platform for sensorimotor learning 5) Evaluation of usability and user-friendliness of web platform for research 6) Evaluation of educational scenarios			
Methodology	The use case will consist of different steps. One is related to the capture of dance motions. These motions will be annotated by experts, and the dance figure recognition module will be tested based on these reference annotations. In a second step, the learning application will be tested by presenting it to new naïve users. The goal of this use case is to design an application which could be displayed in museums/exhibitions/dance festivals/any other space, in order to give a brief introduction to these dances by learning one or a few basic steps, representative of these dances.			
Data Collection Tools	4) Using metrics defined for Body Recognition Module assessment (Objectives 1, 2) 5) Using Questionnaires (Objectives 3, 4, 5, 6) 6) Use of data analytics collected by the final system during the Demonstration phase			
Type of Actors (Number)	1-3 experts and 5-10 learners			
Requirements Examined (D2.1)	(R3C, R3D, R3E, R7(A-E), R8A, R9A, R10A, R11A,R11C, R12E, R13A,R13B, R16A, R17A, R18(A,B),R19(A-B), R20A, R21A, R22(A,B), R23(A,B), R(28-36), R(37-44)			

Table 5-7. Overview of case study "Traditional Walloon Dancing"

Case Study 3	ID	RDI3	Title	Contemporary Dancing
Objectives	1) Assessment of recognition accuracy of dance figures 2) Assessment of accuracy of the identification of different dance styles 3) Assessment of the metric and rules used to evaluate dancers performance 4) Evaluation of usability and user-friendliness of platform for sensorimotor learning 5) Evaluation of usability and user-friendliness of web platform for research 6) Evaluation of educational scenarios			
Methodology	The use case will consist of different steps. One is related to the capture of dance motions from expert contemporary dancers using the optical motion capture system. Several dancers performing the same choreography will be recorded. These motions will be annotated, and the dance figure recognition module will be tested based on these reference annotations. In a second step, the learning application will be tested by presenting it to new naïve users. The learning scenario proposed aims at introducing contemporary dance to young children, which can explore the range of body motion and learn some basic figures in their home.			
Data Collection Tools	1) Using metrics defined for Body Recognition Module assessment (Objectives 1, 2) 2) Using Questionnaires (Objectives 3, 4, 5, 6) 3) Use of data analytics collected by the final system during the Demonstration phase			
Type of Actors (Number)	1-3 experts and 5-10 learners			
Requirements Examined (D2.1)	(R3C, R3D, R3E, R7(A-E), R8A, R9A, R10A, R11A,R11C, R12E, R13A,R13B, R15A, R16A, R17A, R18(A,B),R19(A-B), R20A, R21A, R22(A,B), R23(A,B), R(28-36), R(37-44)			

Table 5-8. Overview of case study "Contemporary Dancing"

5.3.2.3 Traditional Craftsmanship

Case Study 1	ID	TC	Title	Traditional Craftsmanship
Objectives	1) Assessment of recognition accuracy of different wheel-throwing phases 2) Assessment of accuracy of the identification of movements for making different objects 3) Assessment of the metric and rules used to evaluate potter			

	performance 4) Evaluation of usability and user-friendliness of platform for sensorimotor learning 5) Evaluation of usability and user-friendliness of web platform for research 6) Evaluation of educational scenarios
Methodology	The use case will consist of two parts. The first part will consist of capturing of a potter during the wheel-throwing process. These motion capture data will be subsequently annotated by the experts, dividing it onto the different phases. The recognition module will be tested based on these reference annotations as to the accuracy of the detection of different wheel throwing phases as well as the correctness of their execution. There will be at least one test per different object. In a second step, the learning application will be tested by presenting it to new naïve users. The goal of this use case is to design an application which could be displayed in museums/exhibitions, in order to give a brief introduction to wheel throwing pottery dexterities.
Data Collection Tools	1) Using metrics defined for Body Motion and Gesture Recognition Module assessment (Objectives 1,2) 2) Using Questionnaires (Objectives 3, 4,5,6) 3) Use of data analytics collected by the final system during the Demonstration phase
Type of Actors (Number)	1-3 experts and 5-10 learners
Requirements Examined (D2.1)	R3D, R3E, R3F, R8B,R8C, R10A, R11(B-C), R17A, R17E, R18(A,B), R20(B,C), R21B, R22(A,B), R23(A,B), R(28-36), R(37-44)

Table 5-9. Overview of case study "Traditional Craftsmanship"

5.3.2.4 Contemporary Music Composition

Case Study 1	**ID**	CMC	**Title**	Contemporary Music Composition
Objectives	1) Assessment of recognition accuracy of hand gestures 2) Assessment of recognition accuracy of emotional status 3) Evaluation of quality of the produced music 4) Assessment of the metric and rules used to evaluate composers performance 5) Evaluation of usability and user-friendliness of platform for sensorimotor learning 6) Evaluation of usability and user-friendliness of web platform for research			

	7) Evaluation of educational scenarios
Methodology	The use case will consist of different steps. One is related to the capture of musical gestures from expert contemporary music performers using the depth camera, inertial sensors, and EEG capture system. Several pianists (Russian & European school) will perform the same musical pieces (i.e. Beethoven's sonata). These gestures and emotions will be annotated and the Body and Gesture Recognition and Electroencephalography Analysis modules will be tested based on these reference annotations. In a second step, the learning application will be tested by presenting it to new naïve users (both beginners and experienced musicians/composers). The learning scenario proposed aims at introducing contemporary music composition to musicians, who would like to use the intangible musical instrument as well to be able to compose contemporary music.
Data Collection Tools	1) Using metrics defined for Body Recognition Module, Electroencephalography Analysis and Sound Processing Module assessment (Objectives 1, 2) 2) Using Questionnaires (Objectives 3, 4, 5) 3) Use of data analytics collected by the final system during the Demonstration phase
Type of Actors (Number)	2-5 experts and 10-40 learners
Requirements Examined (D2.1)	R1B, R3L, R8G, R10A, R11C, R12A, R17(B,D), R18A, R22(A,B), R23(A,B), R24B, R25(C,D), R27A, R29A, R30(A-E), R31A, R33A, R34B, R35A, R36(A-C), R37(A,B), R38A, R39E, R40(A,C,E), R41(A,D,E), R42A

Table 5-10. Overview of case study "Contemporary Music Composition"

6. Overall Assessment - Evaluation of the Integrated Platform

6.1 Definition of Reference Cases

The overall assessment - evaluation of the integrated platform and its functionalities requires to accord accurate and effectual information that will disclose the actual functionality of i-Treasures platform. To this end, clear reference cases have to be defined in order to perform comparative data analysis between these (control cases) and real cases.

The reference cases will reflect both technical performance and user satisfaction. In this direction, the reference cases will use the *ISO/IEC 25010* standard as a stepping stone for the quality assurance of the platform based on non-functional criteria. The reference cases regarding the technical performance will be developed separately for each module according to use case and sub-use case. They will entail general and commonly eligible performance thresholds for significant technical performance indices (derived from those described in Section 4) These performance thresholds will reflect the minimum admissible performance the integrated platform should exhibit in order to be considered meaningful and functional. Their definition will be realized by the software developers who will rely on current corresponding bibliography.

The second part of reference cases that relates to user satisfaction will concern the overall functionality and usability of the platform as perceived by the user. These reference user satisfaction cases will be defined separately for each sub-use case by the corresponding experts. As in the technical performance reference cases, user satisfaction thresholds will be introduced, based on the user requirements, for all the user satisfaction indices derived from the case studies evaluation phases. These thresholds will represent the minimum and ideal usability conditions of the platform.

6.2 Definition of General Performance Indices

For the overall assessment - evaluation of the i-Treasures platform, general performance indices (GPIs) will be estimated for each use case, i.e., Rare Singing, Rare Dance Interactions, Traditional Craftsmanship, and Contemporary Music Composition, as well as for their sub-use cases (Table 6-1). Additionally, the GPI of the integrated platform will be estimated which will reflect its overall quality. GPIs will be quantitative with values ranging from 0 to 10. The estimation process of the GPIs will be based on a multi-level fusion of technical assessment indices (Section 4), as well as user satisfaction indices (USIs) arising from the case studies (Section 5).

Fusion of indices will be mainly realized using fuzzy inference systems (FISs). A FIS is a system that uses fuzzy logic sets to map inputs (technical assessment and usability evaluation indices) to outputs (more general performance indices). Here, the Sugeno-type FIS [15] will be adopted.

A Sugeno-type FIS with inputs x and y, uses a set of rules to map inputs to the output z. A typical rule in a Sugeno fuzzy model is of the type IF-THEN and it has the form

IF input 1 = x *OR Input 2 =* y*, THEN Output is* $z = ax + by + c$.

The output level z_i of each rule is weighted by the firing strength w_i of the rule that is derived from the corresponding membership functions $F_i(.)$ using a logical

operation. For example, for an *OR* rule with Input 1 = x and Input 2 = y, the firing strength is

$$w_i = OrMethod(F_1(x), F_2(y)), \tag{6.1}$$

where $F_1(x), F_2(y)$ are the membership functions for Inputs 1 and 2, respectively. Usually, the *OR* method corresponds to the maximum of its arguments, thus Equation 6.1 can be written as

$$w_i = \max(F_1(x), F_2(y)). \tag{6.2}$$

The final output of the FIS is the weighted average of all rule outputs, computed as

$$FinalOutput = \frac{\sum_{i=1}^{N} w_i z_i}{\sum_{i=1}^{N} w_i}, \tag{6.3}$$

where N is the number of rules used. Figure 6-1 depicts the aforementioned operation of a Sugeno rule. For more information on the Sugeno-type FIS, the reader is referred to [15].

Here, the membership functions and the rules of the different FISs will be properly selected in order to reflect the significance of the input indices on the output GPI. The different levels of fusion, as well as the performance indices, towards the estimation of the GPI of the integrated platform are described below.

- *Non-Functional Requirements:* A performance index based on the fusion of the qualitative indices (factor analysis) reported in Section 4.1 will be

Rare Singing	**Rare Dance Interactions**		
Canto a Tenore	Tsamiko	**Traditional Craftsmanship**	**Contemporary Music Composition**
Cantu in Paghjella	Calus		
Byzantine Music	Walloon		
Human Beat Box	Contemporary		

Table 6-1. Use cases of the i-Treasures platform (in bold) and their sub-use cases (in the Rare Singing and Rare Dance Interactions cases).

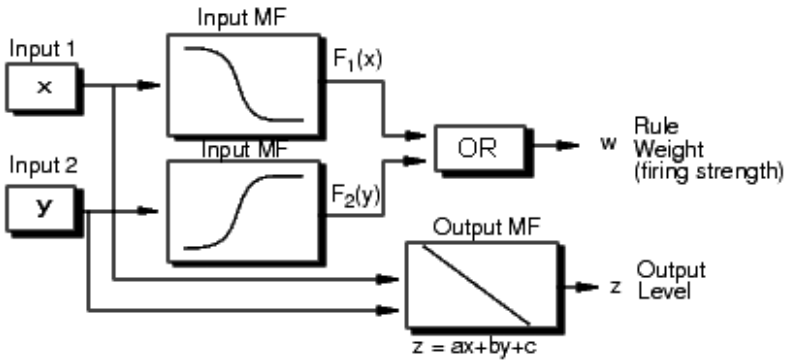

Figure 6-1. Operation of a Sugeno-type fuzzy rule. (Image source: Mathworks Inc.)

produced that will reflect the fulfillment of the non-functional requirements.

- *ICH Capturing and Analysis Modules:* For each of the modules included in Section 4.2, assessment indices will be fused to derive the performance index (PI) that characterizes the quality of the module. Thus, five PIs will be produced corresponding to Facial Expression/Modeling, Body and Gesture Recognition, Electroencephalography Analysis, Vocal Tract Sensing/Modeling, Sound Processing and Text-to-Song modules.
- *Data Fusion and Semantic Analysis:* Indices reported in Section 4.3, will be fuzzified to produce the PI of the data fusion and semantic analysis (FSPI). The FSPI will be estimated separately for each sub-use case of each use case.
- *Educational Process and Platform Interface Assessment:* Indices relating to the technical aspects of the learning management system, the 3D visualisation module and the web platform (Section 4.4) will be fused to produce three PIs, respectively.

For the assessment of the system performance for each sub-use case, appropriate PIs of the ICH capturing and analysis modules, along with the corresponding FSPI will be fused to derive the technical performance index (TPI) for each sub-use case. Where applicable, the PI of the 3D-visualization module will be also used. The TPI and the USI arising from the case study relating to the particular sub-use case will then be fuzzified to produce the GPI of the sub-use case. The average of GPIs for all sub-use cases will form the GPI of the use case. Here, there is no need for a FIS as the GPIs of all sub-use cases are of the same significance. The latter process is illustrated in Figure 6-2(a).

For the assessment of the quality of the integrated platform, the GPIs of all use-cases along with the PIs of the web platform, the learning management system, and the non-functional requirements fulfillment will be fused to produce the GPI of the overall platform, as illustrated in Figure 6-2(b).

6.3 Data Source for Overall Assessment

The overall assessment - evaluation of the integrated i-Treasures platform is going to take place after the completion of the second cycle of case studies, from M40 till the end of the project. The technical data acquired during the second cycle of laboratory testing along with the technical and usability data gathered from the second case study evaluation phase are going to be employed for the overall assessment of the system. Furthermore, informative and significant insight about the system performance and functionality is going to be gained in the course of the system demonstration phase where a large number of potential users are going to experience the i-Treasures platform in physical environment. All the data collection techniques used during the case studies evaluation (questionnaires, software-based user interaction metric analysis, etc.) are going to be engaged for the elicitation of meaningful qualitative and quantitative information from the users who will also play the role of testers. The feedback from the demonstration phase will be valuable since the features and the usability of the system is going to be examined in detail by a variety of users with diverse experience background beyond the tight schedules and strictly predefined scenarios of case studies. The greater the number of the users the more reliable the usability results and conclusions will be.

Consequently, the extended feedback from the demonstration phase along with the technical and usability data collected during the last cycle of laboratory testing and case studies evaluation are going to be used for the thorough assessment-evaluation of the platform and the estimation of the general performance indices of the sub-use

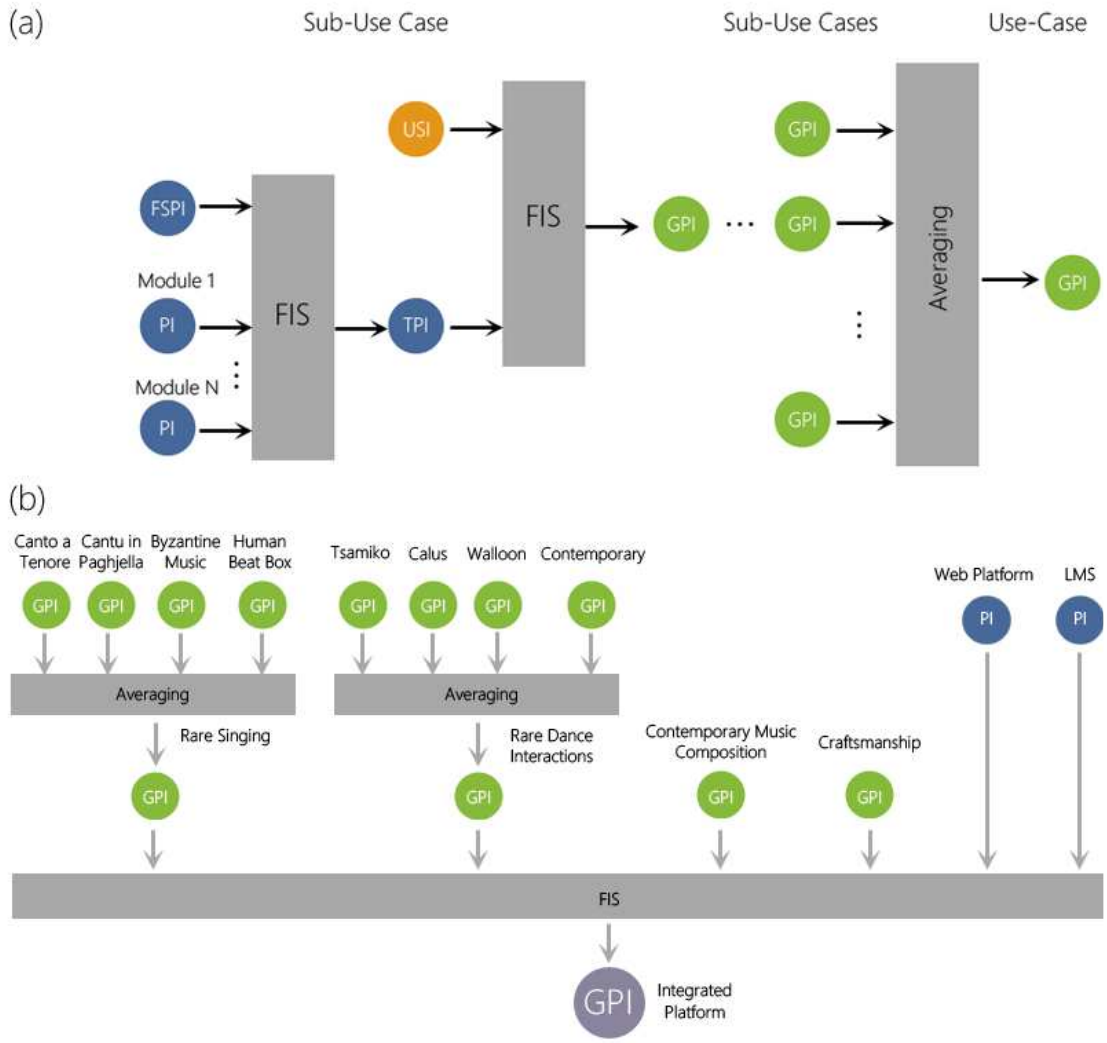

Figure 6-2. (a) Multilevel fusion of performance indices (PIs) and user satisfaction index (USI) for the estimation of the general performance indices (GPIs) of each sub-use case, as well as of the whole use case. (b) Process of the integrated platform GPI estimation.

cases, use cases and the integrated platform according to the aforementioned plan. The general performance indices will reveal and evidence the overall impact (advantages and limitations) of i-Treasures system to potential users and validate the long-term sustainability, utilization and proliferation of i-Treasures as a tool for enhanced education and research practice.

7. Appendix

7.1 Technical Assessment Report Template

Technical Report ID	
Date	
Module/Entity Tested	
Test Leader	

Summary

Summarise what item was tested, what features or combination of features were tested, how the item was tested, what was the approach, what were the main things that happened, what resources were used (tools, people, time)

Variances

If any test items differed from their specifications, describe that. If the testing process didn't go as planned, describe that. Say why things were different

Results

Assessment Indices results

Assessment Category	**Assessment Indices**			
	ID	**Description**	**Desirable Value or Fail/Pass criteria**	**Value**

Other results

Evaluation of Results

How good are the test items? What's the risk that they might fail?

7.2 Case Study Evaluation Report Template

Case Study Report ID	
Date	
Use Case/Sub-Use Case Tested	
Case Study Leader	

Introduction

Provide a brief introduction of the case study

Objectives

Describe the objectives of the case study

Methodology

Describe the case study, how it was conducted, how were the targets evaluated, how were the data collected, what resources were used (tools, people, time)

Variances

In case of unmet objectives describe the reasons that led to this variance.

Results

Describe the results of the case study

Evaluation of Results

Provide comments on the results

8. References

[1] 829-1998 - IEEE Standard for Software Test Documentation, Available at http://standards.ieee.org/findstds/standard/829-1998.html, Retrieved from http://www.cs.otago.ac.nz/cosc345/lecs/lec22/testplan.htm#summary

[2] ISO/IEC 25010:2011, Systems and Software engineering – Systems and software quality requirements and evaluation (SQuaRE) – Systems and software quality models, Retrieved from http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=35733.

[3] Burton-Jonesa, A., Storeya, V. C., & Sug, V. (2005). A semiotic metrics suite for assessing the quality of ontologies. *Data & Knowledge Engineering,* 55(1), 84–102.

[4] Tartir, S., Arpinar, I. B., Moore, M., Sheth, A. P., & Aleman-Meza, B. (2005). OntoQA: Metric-Based Ontology Quality Analysis. *IEEE ICDM Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources*. Houston, Texas.

[5] Gangemi, A., Catenacci, C., & Ciara, M. (2005). A theoretical framework for ontology evaluation and validation. *Proceedings of 2nd Italian Semantic Web Workshop(SWAP)*. Trento, Italy.

[6] MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press.

[7] Laskey, K. (2008). A language for first-order Bayesian knowledge bases. *Artificial Intelligence,* 172(2-3), 140–178.

[8] Manning, C.D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*, Cambridge: Cambridge University Press.

[9] ISO 9241-11:1998, Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability, Retrieved from http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=16883.

[10] ISO/IEC 9126:1991. Information Technology - Software Product Evaluation - Quality Characteristics and Guidelines for the User.

[11] Lewis, C. H. (1982). *Using the "Thinking Aloud" Method In Cognitive Interface Design* (Technical report). IBM. RC-9265.

[12] Ericsson, K. & Simon, H. (1993). *Protocol Analysis: Verbal Reports as Data* (2nd ed.). Boston: MIT Press.

[13] Doubleday, A., Ryan, M., Springett, M., & Sutcliffe, A. (1997). A comparison of usability techniques for evaluating design. *Proceedings of DIS 97*. 101-110. New York.

[14] Runeson, P. & Höst, M. (2008). Guidelines for Conducting and Reporting Case Study Research in Software Engineering. *Empirical Software Engineering*, 14(2): 131-164.

[15] Sugeno, M. Ed. (1985). *Industrial applications of fuzzy control*. Amsterdam, The Netherlands: North-Holland.