SEVENTH FRAMEWORK PROGRAMME FP7

ICT-2011-600545

D6.1 Evaluation methods and plan

Deliverable No.		**D6.1**	
Work package No.	**WP8**	Work package Title	**Project Management**
Authors		**Sally Shalloe, Mirabelle D'Cruz, Natasa Lackovic, Sue Cobb, Charles Crook, (UNott); Craig Alexander, Christopher Chippindale, Giovanna Bellandi, Fred Baker (CAM), Alex Kulik (BUW).**	
Status (F = Final; D = Draft)		**Final**	
File Name:		**3D-PITOTI_ D6.1 Evaluation methods and plan.docx**	
Date		**February 2014**	
Dissemination level (PU = Public; RE = Restricted; CO = Confidential)		**PU**	

Version	Date	Comments
1.0	1st February 2014	Initial draft for comments by PMB and consortium
1.6	18th February 2014	Reviewed document for approval by PMB
2.0	26th February 2014	Document approved by PMB

Keywords	Evaluation, plan, methods, results, analysis
Deliverable leader	Name: Sally Shalloe Partner: UNott Contact: sally.shalloe@nottingham.ac.uk

Table of Contents

List of Tables

Glossary of Terms

ArcTron	Arctron 3D Vermessungstechnik-und Softwareentwicklungs Gmbh, Germany
BUW	Bauhaus-Universität Weimar, Germany
CAM	The Chancellor, Masters And Scholars Of The University Of Cambridge, UK
CCSP	Associazione Centro Camuno Di Studipreistorici Ed Etnologici, Italy
FHSTP	Fachhochschule St. Poelten Forschungs Gmbh, Austria
TUG	Technische Universitaet Graz, Austria
UNott	University of Nottingham, UK
WP	Work Package
UAV	Unmanned aerial vehicle
SfM	Structure-from-motion

Executive Summary

This is the public deliverable D6.1 Evaluation methods and plan, of the FP7 project **3D-PITOTI (ICT-600545)**. This work was carried out as part of WP6 Evaluation and validation, in particular work related to T6.1 Evaluation methods and planning.

The aims of WP6 Evaluation and validation are to identify evaluation methods and define an evaluation plan suitable for assessment of different aspects of the 3D-PITOTI acquisition, processing and presentation systems. Evaluation of the key concepts and technologies will be a continuous process throughout the project. This deliverable report presents the evaluation plan and explains how it has been generated (section 2), describes the evaluation approaches and methods that will be used in the project (section 3), the approach towards generating the evaluation protocols (section 4) and finally the main conclusions (section 5).

1 Introduction

This is the public deliverable D6.1 Evaluation methods and planning, of the FP7 project **3D-PITOTI (ICT-600545)**. This work was carried out as part of Work package (WP) 6 Evaluation and validation, in particular work related to Task 6.1 Evaluation methods and planning.

The aims of WP6 Evaluation and validation are to identify evaluation methods and define an evaluation plan suitable for assessment of different aspects of the 3D-PITOTI acquisition, processing and presentation systems. Evaluation of the key concepts and technologies will be a continuous process throughout the project. The specific objectives of this WP are to:

- Specify a set of methods to evaluate the different hardware and software components, as well as the usability of the 3D-PITOTI presentation system
- Evaluate the effectiveness of the 3D-PITOTI system for scientific analysis of rock-engravings by archaeologists. This will include evaluating the ease of use of the system and its interface as well as usefulness of the manner in which information is presented to scientists, and its potential to enhance scientific understanding of the data.
- Evaluate the usability of the scaled-down version of the 3D-PITOTI Scientists' lab in schools and museums
- Evaluate the user experience for the perspective of the different target end users
- Evaluate the technical components of the 3D-PITOTI system based on a number of assessment criteria including accuracy and reliability
- Evaluate scientific and educational impact of the concept and technologies.

The objectives of Task 6.1 Evaluation methods and planning were to develop an evaluation plan based on the schedule of development. This involved developing the evaluation protocols e.g. the types of users and trials for the specific 3D-PITOTI concepts to be tested, the tasks users will perform using the technologies, collection and analysis of the results; and the development of recommendations. This task included selection of relevant evaluation methods for the formative, summative and technical assessment of the 3D-PITOTI concepts..

Therefore the deliverable is structured in the following way. Section 2 presents the evaluation plan and explains how it has been generated. Section 3 describes the evaluation approaches and methods that will be used in the project. The approach towards generating the evaluation protocols is described in Section 4. Finally the conclusions of this report are discussed in Section 5.

2 Evaluation plan

2.1 Generation of the evaluation plan

The evaluation plan was generated through discussions with the project partners involved both in the development of the 3D-PITOTI acquisition, processing and presentation systems and those who will be end users of the system. The evaluation schedule was based around the timeline of the development of the 3D-PITOTI system technologies and key components and Milestone 13 'Interim evaluation of the 3D-PITOTI system with end users' due in M18 (month 18) of the project. However, as the development of some aspects of the 3D-PITOTI system will be at different development stages, the evaluation plan has taken this into account. During the first 12 months of the project and at the time of this deliverable, a number of deliverables and milestones (D3.1 Ground truth dataset, MS4 Initial SfM 3D reconstruction, MS5 Annotated Ground Truth) were already achieved and evaluated by the project partners with feedback given to the development teams. The evaluation schedule allows for a 2 month gap between the outcome of the

evaluation activity and the next stage of the development (either the interim or final version) to provide time for the feedback from the evaluation activities to be incorporated into the development of the technologies and key components.

Table 2-1: Development and evaluation schedule for main components of the 3D-PITOTI system Table 2-1: Development and evaluation schedule for main components of the 3D-PITOTI system sets out the development and evaluation schedule for the main components of the 3D-PITOTI system.

Table 2-1: Development and evaluation schedule for main components of the 3D-PITOTI system

Component for evaluation	Development and evaluation schedule									
	M12	M16	M18	M20	M22	M24	M28	M30	M34	M36
3D Rock-Art scanner (TUG)		Prototype report D2.2, Interim evaluation activities start	Interim evaluation report				Final evaluation activities start	Final version D2.3		
3D Scanning software (TUG)		Prototype report D2.2, Interim evaluation activities start	Interim evaluation report				Final evaluation activities start	Final version D2.3		
Data post processing Toolkit (ArcTron)		Interim evaluation activities start	Interim evaluation report				Final evaluation activities start	Final version D2.3		
SfM techniques for PITOTI scenes (TUG)		Interim evaluation activities start	Interim evaluation report			Prototype report D3.3	Final evaluation activities start	Final version D3.4		
Automatic acquisition and registration pipeline including final SfM, Viewplanning for optimal SfM acquisition on UAVs and automatic registration software (TUG)			Interim evaluation activities start				Final evaluation activities start	Final version D3.4		
Algorithms for segmentation of 3D PITOTI (FHSTP)	Interim evaluation	Final evaluation activities start	Final version D4.1, MS7							
Multi-user 3D interaction techniques for archaeological exploration of rock-engravings (system for archaeological experts) (BUW)		Interim evaluation activities start	Interim evaluation report			Draft report D5.1.2	Final evaluation activities start	Final version D5.1.3		

Component for evaluation	Development and evaluation schedule									
	M12	**M16**	**M18**	**M20**	**M22**	**M24**	**M28**	**M30**	**M34**	**M36**
Comprehensive database structure following archaeological scientific guidelines incorporating different user interfaces (FHSTP)		Interim evaluation activities start	Interim evaluation report	Database structure MS9					Final evaluation activities start	Final version D4.2
Multi-user multi-touch 3D table-top display (BUW)		Interim evaluation activities start	Interim evaluation report		Final evaluation activities start	Final version MS11				
Appearance preserving real-time rendering techniques (BUW)		Interim evaluation activities start	Interim evaluation report		Final evaluation activities start	Final report D4.3				
Pitoti shape recognition and classification algorithms (FHSTP)		Interim evaluation activities start	Interim evaluation report			Draft report D4.4.1			Final evaluation activities start	Final version D4.4.2
3D-PITOTI animated film (CAM)							Interim evaluation activities start		Final evaluation activities start	Final version D5.3

2.2 Evaluation plan

The evaluation plan is set out in Table 2.2. It shows the start date for each session of evaluation activities for each of the main components of the 3D-PITOTI system; along with the proposed user and technical evaluation methods, the participants who will be involved in the user evaluation activities, and the subsequent reporting of the evaluation activities and outcomes. Some of the components of the system are scheduled for completion significantly in advance of the final evaluation reports due in M36. In these instances, the outcomes of the evaluation activities will be reported in the individual deliverable reports on the development of each component and will be cross-referenced in the final evaluation reports, D6.2 End user evaluation report and D6.3 Technical evaluation report, to avoid duplication.

The evaluation plan contains the proposed methods for evaluation but these may change depending on the development progress for each component. For example, some components are not scheduled to have prototypes available for testing at a specific stage of development so different methods may need to be adopted depending on the development reached at the scheduled evaluation time. It may also be necessary to amend the timescale of the evaluation of a particular component for the same reason. The project will aim to involve potential future end users of the elements of the 3D-PITOTI system in evaluation activities wherever possible, and will hold workshops in Valcamonica and at other partner organisations as required and make use of on-line data collection where necessary. Any deviations from the evaluation plan outlined in this report will be detailed in the subsequent evaluation reports.

Table 2-2 Evaluation activities and reporting schedule

Component for evaluation	Version for evaluation	Start date evaluation activities	User evaluation methods	Participants	Technical evaluation methods	Deliverable	Date(s) of report(s)
3D Rock-Art scanner (TUG)	Prototype	M16	On-site scanning, qualitative evaluation of scanner features using focus groups and questionnaires	Archaeological experts from CCSP, CAM, TUG/ArcTron team members	Component and full system quantitative evaluation of image quality and reconstruction accuracy/precision in the lab/on-site vs. ground-truth	MS13	M18
	Final	M26	On-site scanning, qualitative evaluation of scanner features using focus groups and questionnaires	Archaeological experts from CCSP, CAM, TUG/ArcTron team members	Full system quantitative evaluation of image quality and reconstruction accuracy/precision in the lab/on-site vs. ground-truth	D2.3*D6, D6.3	M30, M36
3D Scanning software (TUG)	Prototype	M16	On-site ease of use and feedback quality using heuristics, focus groups and questionnaires	Archaeological experts from CCSP, CAM, TUG/ArcTron team members	Evaluation of stability, synchronization behaviour, CPU load/battery runtime	MS13	M18
	Final	M26	On-site ease of use and feedback quality using heuristics, focus groups and questionnaires	Archaeological experts from CCSP, CAM, TUG/ArcTron team members	Evaluation of stability, synchronization behaviour, CPU load/battery runtime	D2.3*, D6.2, D6.3	M30, M36
Data post processing Toolkit (ArcTron)	Interim	M16	In-house user trial to verify state of art of processing algorithms	ArcTron staff, TUG, BUW, FHSTP,	Component testing as far as several algorithms are developed	MS13	M18
	Final	M26	User trial of developed algorithms and their improvement	ArcTron staff, TUG, BUW, FHSTP, CAM, Archaeocamuni	Full system testing by ArcTron stuff, esp. archaeologists employed at ArcTron; system test by partner companies (if allowed and possible)	D2.3*, D6.2, D6.3	M30, M36
SfM techniques for PITOTI scenes (TUG)	Interim	M12	Not applicable as no user interaction	Not applicable	Quantitative Evaluation in Lab to Ground-Truth	MS13	M18

Component for evaluation	Version for evaluation	Start date evaluation activities	User evaluation methods	Participants	Technical evaluation methods	Deliverable	Date(s) of report(s)
	Final	M16	Not applicable as no user interaction	Not applicable	Quantitative Evaluation in Field to Ground-Truth	D3.4*, D6.2, D6.3	M30, M36
Automatic acquisition and registration pipeline including final SfM, Viewplanning for optimal SfM acquisition on UAVs and automatic registration software (TUG)	Interim	M18	Observations, focus groups and questionnaires	Archaeological experts from CCSP, CAM, TUG/ArcTron team members	Comparison to Ground-Truth Laser scans as acquired in T3.2 in terms of completeness and accuracy	D6.2	M36
	Final	M26	Observations, focus groups and questionnaires	Archaeological experts from CCSP, CAM, TUG/ArcTron team members	Comparison to Ground-Truth Laser scans as acquired in T3.2 in terms of completeness and accuracy	D6.2, D6.3	M36
Algorithms for segmentation of 3D PITOTI (FHSTP)	Interim	M12	None, no user interface	TUG, FHSTP	Quantitative evaluation against ground truth, component testing	D4.1*, D6.2, D6.3	M18, M36
	Final	M16	None, no user interface	TUG, FHSTP	Quantitative evaluation against ground truth, full method testing	D4.1*, MS13	M18
Multi-user 3D interaction techniques for archaeological exploration of rock-engravings (system for archaeological experts) (BUW)	Interim	M16	Expert walkthrough, Semi-structured interviews, Focus group	Archaeological experts from CCSP, CAM & naïve users, e.g. students of architecture and urban planning	none	MS13	M18
	Final	M26	Expert walkthrough, Semi-structured interviews, Focus group, Task performance measures		none	MS8,D5.1.3, D6.2, D6.3	M30, M36
Multi-user multi-touch 3D table-top display (BUW)	Interim	M16	Expert walkthrough, Semi-structured interviews, Focus group	Education experts from CCSP, UNott & naïve users, e.g. students of architecture and design.	none	D6.2, D6.3	M36
	Final	M22	Ethnographic methods	Teachers, museum Curators and visitors	none	MS11, D5.1.2*, D6.2, D6.3	M24, M36

Component for evaluation	Version for evaluation	Start date evaluation activities	User evaluation methods	Participants	Technical evaluation methods	Deliverable	Date(s) of report(s)
Appearance preserving real-time rendering techniques (BUW)	Interim	M16	Perceived difference between images, perceived changes	Archaeological experts from CCSP, CAM & naïve users, e.g. students of architecture and design	Absolute difference between images from different data representations	MS13	M18
	Final	M22				D4.3*D6.2,D6.3	M24, M36
Pitoti shape recognition and classification algorithms (FHSTP)	Interim	M16	None, no user interface	None	Quantitative evaluation against ground truth, component testing	MS13	M18
	Final	M32	None, no user interface	None	Quantitative evaluation against ground truth, full system testing	D6.2, D6.3	M36
Pitoti database with structure following archaeological scientific guidelines incorporating different user interfaces (FHSTP)	Interim	M16	Interviews	Experts from Arctron, FHSTP, CAM, CCSP	Component testing	MS13	M18
	Final	M32	Interviews	Experts from Arctron, FHSTP, CAM, CCSP	Full system testing	D6.2, D6.3	M36
3D-PITOTI animated film (CAM)	Interim	M26	Test screening in a heritage, schools and academic context. Questionnaire	Project partners, School children and adults	Quantitative and qualitative questionnaires		
	Final	M32	Film screening in a heritage, schools, and academic context. Questionnaire	Public		D5.3	M36

*This will be cross-referenced in D6.2 & D6.3

3 Evaluation approaches and methods

There are three types of evaluation that are proposed in this evaluation plan, formative, summative and technical evaluations. The following definition by Trochim (2006) explains the first two which will be undertaken during T6.2 Formative and summative evaluation.

"Formative evaluations are used to strengthen or improve the object being evaluated -- they help form it by examining the delivery of the program or technology, the quality of its implementation, and the assessment of the organizational context, personnel, procedures, inputs, and so on. Summative evaluations, in contrast, examine the effects or outcomes of some object -- they summarize it by describing what happens subsequent to delivery of the program or technology; assessing whether the object can be said to have caused the outcome; determining the overall impact of the causal factor beyond only the immediate target outcomes; and, estimating the relative costs associated with the object." (Trochim, 2006).

Formative evaluation will take place during the development phase of the components of the 3D-PITOTI system and the outcomes from the evaluations will be used to inform the subsequent development phase. Summative evaluations will be carried out toward the end of the development phase to establish whether the component has met the needs of the end users of the system, and whether it enhances the acquisition and processing of data as well as the interpretation and sharing of information, ideas and challenges.

Technical evaluation, which will take place in T3.6, will comprise of a functional component and full system testing of the 3D Rock-Art scanner and UAV system. These tests will assess issues such as robustness, accuracy, reliability, safety, etc. Initial testing will take place within laboratory settings followed by on-site testing in different conditions. The time, effort and required expertise will also be documented and the technical outputs will be compared to the initial first high-quality scans from T3.2 Ground-Truth Data Acquisition, reported in D3.1 Ground Truth Dataset.

The methods to be used during the evaluation activities will address user and organisational requirements as well as usability issues and educational impact, as these will ultimately be used to judge the success of the tools. The requirements specification set out in D1.2 Specification of 3D-PITOTI system is essential so that the attributes can be measured and tested. The attributes may include functional (i.e. what the tool needs to be able to do) and non-functional requirements e.g. understandability (how easily can a user understand what the tool can be used for); learnability (how easily can a user learn to use the tool); operability (how easily can a user use the tool); attractiveness (the capability of the tool to be liked by the user); acceptability; efficiency and satisfaction.

A number of evaluation methods will be used to support the different aspects of the 3D-PITOTI system in terms of the types of evaluations planned and the resources available. The methods described in section 3.1 are proven techniques within the field of ergonomics for analysing and improving user interactions with technologies and systems. These methods include: questionnaires, interviews, workshops, focus groups, observations, cognitive walkthrough, heuristics and guidelines. Some of these methods are also widely used in educational evaluation (Charles, 1998) and user experience evaluation for cultural heritage (Gockel et al, 2013). For the 3D-PITOTI scientists' lab being developed, specific tools may be used for the evaluation of aspects of virtual reality systems including input and output devices and interfaces along with general usability issues.

3.1 General evaluation methods

Depending upon the circumstances, different methods may be used to collect and analyse data, either singularly or in combination. The choice of method will be based on a variety of factors including: the object or system to be evaluated, the environment and context in which the evaluation is to be conducted, the time and resources available, access to sites and the personnel and expertise of the evaluation team.

3.1.1 Questionnaires

Questionnaires are a widely used method of gathering information from users. They can be administered in either face to face, by telephone, online or on paper. Questionnaires are often used to gather relatively small amounts of data from a large number of participants. They are quick to administer and can be relatively quick to analyse (Robson, 2002). They can be made up of open or closed questions or a combination of the two (Rogers, Sharp & Preece, 2011). Questionnaires can be a sole method of gathering data or may be used in conjunction with other methods (Rogers, Sharp & Preece, 2011)and can be used to gather both qualitative and quantitative data (Bowman, Gabbard & Hix, 2002).

Questionnaires may be specifically designed for the purpose of a study but there are also standard, well established questionnaire for assessing particular attributes. Relevant examples for post study evaluation to look as aspects such as efficiency, control and learnability include System Usability Scale (SUS) developed by John Brooke in 1986 (http://www.usabilitynet.org/trump/documents/Suschapt.doc) and System Usability Measurement Inventory ,SUMI (http://sumi.ucc.ie/), used for measuring software quality from the end user's point of view (http://sumi.ucc.ie/whatis.html). Another example, the Usability Metric for User Experience (UMUX) is a four-item scale used to assess the subjective usability of an application (Finstad, 2010). Others can be used to assess aspects such as cognitive workload e.g. the Subjective Mental Effort Questionnaire (SMEQ) (Sauro & Dumas, 2009).

3.1.2 Interviews

Interviews typically involve a researcher asking participants to directly respond to questions. This can be carried out in person, on the telephone or via other means of communication e.g. video calling. However, interviews always involve real-time communication between the researcher and the respondent (Robson, 2002; Rogers, Sharp & Preece, 2011). They are good for obtaining subjective opinions and reactions and often go into more detail than questionnaires (Bowman, Gabbard & Hix, 2002).

The different types of interview are often categorised into structured, semi-structured and unstructured interviews. These categories can be linked to the 'depth' of information needed to be collected. A highly structured interview is really a questionnaire that is delivered in person with fixed questions in a pre-determined order. In contrast an unstructured interview is guided by a broad topic but the respondent is free to say what they like on the topic with little prompting by the interviewer. In a semi-structured interview the interviewer will have prepared a few broad questions in advance of the interview, but will ask additional unplanned questions during the interview to prompt further information from the respondent. This allows the researcher follow interesting lines of enquiry in further depth. When it is clear the respondent has no more to say about a topic, the researcher then moves on to the next.

A semi structured or unstructured research interviews are most appropriate when:

- A study focuses on the meaning of a particular phenomenon to the participants
- There is a need want to find out about individual perceptions of a process/object
- Where historical accounts are required
- For exploratory information gathering
- When quantitative information has been gathered and there is a need to validate particular findings or illustrate the meanings of new findings (adapted from Robson, 2002)

Unstructured interviews are often used as exploratory interviews where they are used to gain ideas rather than data. They are less formalised and structured around a 'hidden agenda' which is a list of topics and questions that the interview wants to cover but the order in which they are approached is set by the respondent. The quality of information gained from this method is heavily depended on the skills of the interviewer and the respondents' willingness to talk openly and honestly. Less structured interviews allow for greater interviewer and respondent flexibility (e.g. the ability to investigate certain aspects in more depth) and can provide more insight into participants' thought processes (Bowman, Gabbard & Hix, 2002; Robson, 2002).

Advantages of interviews include:

- They are flexible
- Non-verbal messages can give insight
- Provide rich data
- Good response rates

Disadvantages include

- Lack of standardisation (depending on type)
- Bias may be present
- Time consuming
- Can be expensive (Rogers, Sharp & Preece, 2011)

Interviews can be used as the sole method or in conjunction with others. For example, following a more formal experimental situation, an interview may be used to gain information on the respondent's thoughts on the subject matter or to complement observations made in a user trial (Robson, 2002). In addition, aids such as demonstrations may be used alongside interviews to assist the participant in providing their responses (Bowman, Gabbard & Hix, 2002). The results obtained are typically qualitative although some methods of analysis which can be used (e.g. Theme Based Content Analysis (Neale & Nichols, 2001)) will provide quantitative results from qualitative data. It is also possible to obtain quantitative data from structured interviews (Rogers, Sharp & Preece, 2011).

3.1.3 Focus Groups

A focus group interview is a group interview on a particular topic - the 'focus'. It is an open-ended group discussion which is guided by the researcher and usually lasts an hour or more. It can also involve other activities such as watching videos, using prototypes, having a demonstration and other activities. The size of the group varies but the optimum size is felt to be between 8 and 12. Focus groups can be used as the primary data collection method in studies but are commonly used with other methods such as observation and individual interviews. Having a number of participants present at once can be useful for generating discussions and debate about specific issues. This is often particularly useful in the early stages of processes (e.g. for idea generation) and when assessing prototypes (Robson, 2002). They are also used to help develop more structured methods such as questionnaires or alternatively, used to gain more information about particular responses on questionnaires.

The researcher guiding the focus group (referred to as the moderator) has to both regulate the group and also facilitate it, along with striking a balance between taking an active and passive role. The moderator has to generate interest in the topic but not lead the discussion in any preconceived direction whilst making sure no one person dominates the discussion and that rules of politeness and turn taking are adhered to. It is useful to have another researcher present to make notes on who is speaking (as when transcribing audio tapes it may be difficult to work this out) and also to give feedback on the moderator's performance.

When analysing data from focus group interviews, care must be taken to consider group dynamic issues. For example, if people keep quiet it could signify that they agree with the point being made but it could also signify that they were unwilling to voice their objections (Robson, 2002).

3.1.4 Observations

Observational methods are used to collect information about human performance relating to a huge range of activities. Watching people carrying out a task can provide data on performance time, errors, task sequences and worker interaction as well as providing some insight into the difficulty or ease of carrying out a task. Observational methods can be broadly split into two types: direct and indirect. Direct observation (and recording) of behaviour involves the researcher being present during the task either in the immediate location or using remote observation techniques such as closed-circuit television. Indirect observations are taken where the task is viewed at another time from when it took place by other means such as video recording or time-lapse photography i.e. after the event. When carrying out direct observation, care must be taken to consider and control for observer effects (Fostervold et al, 2001).

A disadvantage of direct observation when the observer is in the immediate location is that their presence can have an effect on the task performance (Robson, 2002). Indirect observation involves a researcher observing the task after it has been carried out. This may be done through the use of video recording (Rogers, Sharp & Preece, 2011).

Observation can be either structured or unstructured. Unstructured observation involves the researcher noting anything that is of interest. Structured observation involves the observer noting instances of predefined events, behaviours or themes (Robson, 2002). Observation is a time consuming method both in terms of conducting observations and analysis (Robson, 2002). However, large amounts of rich data can be obtained. This data can be both qualitative and quantitative in nature (Rogers, Sharp & Preece, 2011). Observation can be used in conjunction with other methods to complement the data obtained from these. Video analysis is one such tool which may be used to support observations.

3.1.5 Heuristics and Guidelines

Heuristics or guideline based evaluation is a method of expert appraisal. Typically, several experts will evaluate a system separately against a set of relevant heuristics or guidelines. The results of the individual analyses are then amalgamated and the resulting issues are prioritised (Bowman, Gabbard & Hix, 2002; Hix & Gabbard, 2002; Nielsen, 1994). Some propose that a proposed suitable number of evaluators is three to five as this should enable a large proportion of issues to be identified without too high an expense (Nielsen & Mack, 1994). However, other studies have shown that having more evaluators is better but is more expensive (Rogers, Sharp & Preece, 2011).

Heuristic and guidelines based evaluation does not make use of any representative users (Bowman, Gabbard & Hix, 2002). It is therefore a relatively quick method but generally does not reveal as many issues as other methods which do make use of end users. Heuristics and guideline based evaluations are often most useful early on in the development process as they will reveal obvious issues before user testing therefore having the potential to reduce the time and cost of the evaluation process (Bowman, Gabbard & Hix, 2002; Nielsen & Mack, 1994).

3.2 Technical evaluation approaches

In the project it is expected that each partner with responsibility for developing technologies will be in charge of their (i) Quality Assurance policy (e.g. unit/integration tests), and (ii) of applying the agreed system-level tests to their component. As the technical evaluation and validation is integrated with the system development procedures and therefore varies for each component, the approaches to technical

verification and validation are presented separately. Despite these differences, common themes can be identified throughout the process.

Software and hardware testing is often split into two parts, verification and validation, where verification tests the product of a development phase and checks it against the targets set at the beginning of that phase, and validation tests the finished system against the requirements specification. Computer software is often developed on the basis of assumptions with elements of ambiguity; this makes it prone to defects. It can combine aspects of machine, maths, language and thought. This means that, whilst it can be very powerful, it can also be unpredictable and therefore risky. Therefore technical testing is the art of uncovering the unknown and therefore can be difficult to plan. As with user evaluation methods, testing is more efficient if there is a well written requirements specification that is clear, consistent, unambiguous, measurable and testable. This requirements specification should also state what the software should not do where appropriate. The requirements for the 3D-PITOTI system will be set out in D1.2 Specification of the 3D-PITOTI system and D2.1 Requirements of 3D-PITOTI scanner. Another way to make the correction of software defects easier is to have a rigorous version management system. This will make it easier to track down problems as well as give users precise information about the issues with specific releases.

3.2.1 Types of technical testing

To avoid the build up of defects and the consequent increasing difficulty of diagnosis, testing should take place early and often. For example, this means testing individual modules or units to see how they perform rather than waiting for them to be integrated into the system. This unit testing can then be followed by integration testing, where the units are linked to others to see if they work in combination. A part of this is regression testing which can be used to check whether changes in one unit result in the incorrect operation of another. Finally, system testing, or validation, can be carried out to find out if the finished hardware and software conforms to the specification. Within each partner organisation development and testing will be carried out separately to ensure that the testing is independent from development bias or preconceptions. The stages of development of test plans are detailed below.

Stage 1: what to test?
In the first instance for each release of a component (as shown in Table 2), testing will be done by the developer to find out if it is broadly is working as expected. Once this has been done it will be handed over to testers who will then plan out how the verification should be carried out. The first stage of this will be to form a list of the elements that make up the unit or system. These might include – functionality, code structure, user interface, internal interfaces, interfaces to other systems, input ranges, outputs, manuals and other physical components, data structures, platform & environment, and configuration elements. For each of the relevant elements it will then be necessary to decide what aspects of that element to test. Some elements may require more than one test; some tests may test more than one element.

Stage 2: how to test?
The next step will be to decide how each test will be carried out. Most elements can be tested twice – once to see if they give the correct response to valid inputs (positive testing), and again to see that they give an appropriate response to invalid inputs (negative testing). It is possible that there could be an infinite variety of invalid inputs so negative testing may have to be limited to the more likely of these. To document this process, for each test a test case will be written. This is not only used to record the tests carried out but also to provide accurate information to the developer in the event of a defect being found. A test case might include the following items – unit or module description, priority, initial item configuration, software/hardware configuration, test steps, expected behaviour, actual behaviour/test outcome. For each of these items the tester will provide all the relevant details of the tests to be performed. At this stage the last item will be left blank.

Stage 3: prioritising

The resulting list will now include all the tests that have been considered. However, some of these tests may not be viable as the cost of, or time to carry them out may not be justifiable in terms of the nature of the defects that might be found. To work out which tests will be performed, the next stage will be to prioritise them with respect to risk. In the first instance this will be done by asking the developers where they think there might be defects. For each test it will be necessary to establish the likelihood of a problem and the severity of the effects of the problem on the correct function of the component. Having prioritised the tests an estimate will be made of the cost of the test (probably in terms of time taken to perform the test). Now low priority, high cost tests will be eliminated from the list of tests to be performed.

It is likely that there will be more than one phase of testing, as defects are found, corrected and the system retested. If no defects are found, the testing plan will be revisited to see whether it is missing something. The purpose of testing is to find problems, if no problems are found the testing has probably failed. It is very rare for software in particular to be completely bug-free.

3.3 Educational evaluation approaches

3.3.1 General approaches

Educational evaluation approaches can also be distinguished as 'formative' and 'summative'. The former evaluates the learning as it develops and throughout some course or educational experience with emphasis on furnishing ongoing learning feedback to optimise the direction and pace of change. Formative assessment methods involve collecting outputs of learners' engagement during the course or educational experience. Summative evaluation involves evaluating learning at the end of a course of study or some learning cycle with methods commonly involving tests and/or assignments. In all cases there is a need to demonstrate reliability and validity: creating confidence in the reproducibility of results and confidence about the authenticity of what has been measured.

By the end of a course or educational experience, learners are expected to have developed understandings that accord to educational goals and objectives. Stufflebeam (2001) provides an overview of evaluation approaches in educational courses, some of which are applicable for shorter-term educational initiatives and events and thus appropriate for educational evaluation within the project. The goals and objectives can be formulated as a subject-related list of what students will be able to comprehend and/or do, or what capabilities and skills they will have developed by the end of the educational event. The goals and objectives are formed under the influence of the educators (e.g. through day to day planning), as they act within institution, local, national and international educational systems.

We have extracted and adapted eight approaches to educational evaluation (out of 22 listed by Stufflebeam, 2001) as they can be related to educational initiatives and interventions while the remaining approaches are more focused on course evaluation as a mechanism of institutional policy and strategy (Stufflebeam, 2001, p. 16-80). Those eight approaches are:

1) *Question-based evaluation:* This develops and applies specific evaluation questions that are given to learners. The questions can be applied on educational premises or distributed online. This can be done using a variety of methods (e.g., surveys, structured interviews, semi-structured interviews, focus groups). The questions are usually narrowly defined and may reflect behavioural or operational objectives, or an expert's preferred set of criteria.

2) *Standardized tests*: These are a particular type of question-based evaluation. They are regularly administered tests, standardised for large-scale examination, often consisting of multiple choice questions or otherwise structured formats.

3) *Performance testing*: This approach is oriented towards students' performance and what they can do and create as a sign of their learning. It evaluates learning as a product of students' engagement individually or in a group, requiring "students to demonstrate their achievements by producing authentic responses to evaluation tasks, such as written or spoken answers, presentations, portfolios of work products, or group solutions to defined problems (Stufflebeam, 2001, p.25)."

4) *Experimental studies*: This evaluation approach suggests using structured experiments with experimental and control groups – the first receiving an intervention, the latter not - and then contrasting the outcomes.

5) *Case study evaluation*: This evaluation approach is a focused, in-depth description and analysis of the learning activities, actors, environment and various contextual factors surrounding it without controlling those circumstances in any way.

6) *Criticism and Connoisseurship*: The purpose of this evaluation approach is to describe, critically appraise, and illuminate a particular course or event's merits. The evaluation questions are defined by expert evaluators. Possible questions are: "What are the (educational) course/initiative's essence and salient characteristics?" "What merits and demerits distinguish the particular course/initiative from others of the same general kind?"

7) *Stakeholder–centred evaluation approach*: This approach evaluates the needs, experience and responses of those who deliver courses and those who attend them along with other relevant stakeholders, for example: learners, parents, teachers, schools, university professionals. It can be concerned with stakeholder's affective responses.

8) *Design-based approach*: This evaluation approach evaluates an educational design and/or resource for teaching and learning over a period of time and in stages. It can apply a variety of methods: observation, experimental intervention, interviews, artefacts and discussion record (for an analysis of created artefacts, spoken and written engagement).

Evaluation methods are commonly grouped according to whether they generate qualitative or quantitative data. Using a mixed-method evaluation approach has become more and more common.

Typical methods generating quantitative data in educational research include:

- Questionnaires
- Structured interviews
- Standardised testing (annual)
- Experiments
- Rating scales (e.g. Likert)

Typical methods generating qualitative data in educational research include:

- Text , artefact and/or discussion recording (for performing content /multimodal/ critical discourse/ theme/corpus linguistics analysis)
- In-depth descriptions
- Semi-structured and open-ended interviews
- Focus groups

- Photo elicitation (participants are collecting or shown photographs to provide reflections on an issue)
- Independent and participant observations

It is important to note that with the proliferation of multimodal digital resources and multimedia objects and forms of communication, many authors argue in favour of going beyond language-focused testing, allowing for different hybrid forms to legitimately be explored and introduced as evaluation resources (e.g. students' multimedia productions and expressions).

3.3.2 Evaluating the impact of a resource on learning

At its simplest, the phrase "educational evaluation" suggests exploring a relationship between two things: a learning intervention (e.g., a digital resource) and a learning outcome (typically, learner knowledge). Judging whether a resource (e.g. a 3D environment) has an 'impact' on learning suggests a comparison: that is, outcomes observed *with* the resource versus outcomes observed *without*. Yet this simple conception is often dismissed for being a "medical model" of educational evaluation. Sceptics argue that an educational intervention rarely follows the model of injecting a circumscribed drug into a passive and receptive organism (Clark, 1994).

An educational resource is better considered a 'tool' rather than a pharmacological messenger. As a tool, it will be exercised in different ways: variation dictated by different teachers, different learners, and different learning contexts. We can think of this inevitable teacher-learner-context mix as an 'activity system' – a framework of activity that our intervention disturbs, by inviting it to assimilate some new tool or resource.

This means that what is observed in an evaluation is not a 'resource' but a 'resource used that way'. If specifying our resource (input) is therefore not simple, neither is specifying our learning (output). When the learner leaves this activity system they may well be changed – entering some new activity system equipped with new knowledge. They have learned. But that knowledge may be anchored to the circumstances of the first (learning) activity system: put simply, it may be more or less flexible; the learner may be more or less versatile in applying what has become known across a whole range of contexts.

An educational evaluation needs to consider all of this. In practice this is likely to mean that "evaluating" has two main foci. Firstly, we may evaluate the variety of ways in which a novel resource is assimilated into human activity systems – how a resource affords differing modes of educational practice, all the ways the tool may be used. Ideally this would be pursued through the methods of "design-based (educational) research" (Amiel and Reeves, 2008). Resource users (teachers and learners) are observed or questioned in ways that expose how the properties of the resource invite particular modes of practice: defining an iterative process of re-design. If the teacher herself evaluates a new resource or an educational design in a collaborative relationship with researchers, this is then the case of an 'action research' approach to evaluation (Brydon-Miller, 2003). The methods applied within design-based research and action research depend on what needs to be evaluated and what answers are sought for.

The second focus of evaluation would be on judging changes in the learner. The changes sought might be connotative (what is known), cognitive (the process of knowing) or affective (the emotions and experience of knowing). These may be assessed formatively and summatively.

3.3.3 Evaluating an impact of visualisations on learning

When it comes to visualisations, Naps et al. (2003) provide a set of methods for evaluating the educational impact of visualisations. Evaluation can target teachers or students or both. For example, teachers can evaluate visualisation tools by using Likert scale values (e.g., strongly disagree, disagree, neutral, agree,

strongly agree) and by answering multiple-choice or open-ended questions. The Likert-scale evaluative statements could include 'The tool is easy to show and teach to students' or 'The tool works reliably'. An open-ended question might be 'How did students interact with the tool?' Naps et al (2003) suggest measuring student satisfaction (affective evaluation) using similar statements Likert-scale evaluative statements such as 'I enjoy using the tool', 'I feel I understand the concept better when using the tool' along with open ended questions such as 'How did you use the tool'?

In terms of formative evaluation, the following can be evaluated:

- Students' attention to a visualisation (by observing students and documenting what they do)
- Immediate and intermediate student feedback (by asking and recording questions)
- Students' opinions (by interviews)

In terms of summative evaluation, it is suggested that this may be realised using:

- Pre and post-content tests
- Using Bloom's taxonomy (six "levels" of learners' understanding: knowledge, comprehension, application, analysis, synthesis, and evaluation). This approach is contested if used hierarchically since different levels may not occur in a strict hierarchical order.
- Attitude surveys.

4 Evaluation protocols

When developing the evaluation protocols for the elements of the 3D-PITOTI system, the project will consider what type of users will test the system, what will be tested, the tasks that will be carried out, what data will be collected and how this will be analysed.

4.1 Participant selection

Participant selection for the evaluation of each use element of the system will be based on the task that is being carried out along with the tools and methods being used. For example, where heuristics or guidelines are used, participants can be experts but where questionnaire studies are conducted as part of user evaluation, representative end users of the system will be more applicable.

There are a number of methods of participant sampling which can be used. These can be generalised into two broad categories: probability and non-probability sampling. Probability sampling recruits participants from a target population and there are a number of ways that this can be done. Random and stratified sampling is the most common approaches. Random sampling involves using a random number generator or choosing every n^{th} person from a list. Stratified sampling involves dividing the population into groups and then applying a random sampling technique to each group. Non-probability sampling usually involves participants who are self-selected or recruited based on availability. It can be more difficult to make robust generalisations when using non-probability sampling (Rogers, Sharp & Preece, 2011). However is recognised however that within the constraints of the project, non-probability sampling is likely to be used as the possible population for some elements of the 3D-PITOTI system include the public (as museum visitors) and archaeologists working in specific fields and probability sampling from these populations would be almost impossible. Wherever possible, participants for user trials will be recruited from actual potential user groups.

Sampling size will also be considered when selecting participants. Appropriate sample sizes will depend on the methods being used but for certain studies, minimum sample sizes can be determined using power analysis.

4.2 Selection of elements to be evaluated and tasks to be carried out

The elements that will make up the 3D-PITOTi will be selected for development in the requirements stage and those relating to the visualisation system are specified in D1.2 Specification of the 3D-PITOTI system and those relating to the 3D Rock-Art Scanner are specified in D2.1 Requirements of 3D Pitoti Scanner. The elements are detailed in these reports and appropriate tasks for evaluation will be defined. The evaluation tasks to be carried out will be selected based on the elements being tested and the data desired.

4.3 Data collection and analysis

The methods and tools selected for use in evaluation will be chosen based on their appropriateness. This includes consideration of the type of data that are required and the evaluation questions being asked. In addition, participant characteristics may affect methods and tool selection. Methods and tools do not have to be used in isolation but can be combined as this can result in obtaining richer data and broader understanding.

The complexity of tasks which participants will be asked to carry out and the amount of interaction required may affect the way in which methods and tools are used. For example if a task is complex, or requires a lot of interaction, it may be more practical to carry out observations or conduct post-trial interviews or ask participants to complete post-trial questionnaires. For less complex or less interactive tasks, questionnaires or interviews conducted in-situ may be more suitable. This is particularly the case where the task is not easy to observe. Consideration has also been given to the location in which evaluation will occur. Evaluation can take place in laboratories or in field settings (Rogers, Sharp & Preece, 2011, Egger et al, 2013). Given the nature of 3D-PITOTI systems being developed concepts, studies will take place both in laboratories and in the field.

Practical issues which need to be considered when planning data collection and therefore the methods and tools that will be used include: selecting participants, resource availability and researcher expertise. The availability of resources can affect the tools and methods which are used and can also affect the location in which studies are conducted. In addition, some methods and tools will require specific expertise in order to facilitate the studies or analyse the results (Rogers, Sharp & Preece, 2011).

Methods of data analysis will be selected based on the type of data collected, the methods used and the research questions. Quantitative data collected, such as questionnaire responses or timings of user interactions will often be analysed using appropriate statistical tests. Qualitative data is often more difficult to analyse. Possible approaches include counting word or phrase frequencies and coding data (either with predefined codes or emergent codes from researcher interpretation). Care should be taken when selecting an analysis approach. If there is a need for quantifiable data, counting frequencies of phrases or occurrences may be more appropriate. If the analysis aims to look for specific occurrences, predefined codes may be most appropriate. If the study is more exploratory, emergent codes may be more appropriate.

5 Conclusions

The objectives of Task 6.1 Evaluation methods and planning were to develop an evaluation plan based on the schedule of development. This report presents the evaluation plan which has been developed with the partner organisations responsible for the technical development of the components of the 3D-PITOTI system, the partner organisations who will be among the eventual end users of the system, and those who will be responsible for carrying out the evaluation activities. The schedule for evaluation and reporting has been suggested along with the proposed methods and users to be involved. The relevant evaluation

approaches and methods have been defined and this document will act as a resource for the project partners to select appropriate methods, participants and evaluation metrics for the evaluation protocols.

References

Amiel, T. & Reeves, T.C., (2008). Design-Based Research and Educational Technology: Rethinking Technology and the Research Agenda. *Educational Technology & Society*, 11(4), p.29–40. Available at: [Accessed September 22, 2013].

Bowman, D. A., Gabbard, J. L. & Hix, D. (2002). A Survey of Usability Evaluation in Virtual Environments: Classification and Comparison of Methods. *Presence: Teleoperators and Virtual Environments*, 11(4), 404-424. doi: 10.1162/105474602760204309.

Brydon-Miller, M., Greenwood, D. & Maguire, P., (2003). Why action research? *Action research*, 1(1), 9–28. Available at: [Accessed September 22, 2013].

Charles, C. M.1998. *Introduction to Educational Research. Third Edition*. Reading, MA, Addison Wesley Longman.

Clark, Richard E. (1994). Media will Never Influence Learning. Educational Technology Research and Development, 42(2), 21-29.

Clark, Richard E. (1994). Media will Never Influence Learning. *Educational Technology Research and Development*, 42(2), 21-29.

Hix, D. & Gabbard, J. L. (2002). Usability Engineering of Virtual Environments. In K. M. Stanney (Ed.), *Handbook of Virtual Environments: Design, Implementation and Applications* (pp. 681-699). Mahwah, NJ: Lawrence Erlbaum Associates.

Egger, U., M. Seidl, P. Judmaier, F. Baker, C. Chippindale, M. Grubinger, N. Jax, G. Seidl, and C. Weis, "Multi-touch Rocks: User Experience Metrics for a Multi-user Game on a Multi-touch Table", Technical Report TR 01-2013, St. Pölten, Austria, 2013. http://mc.fhstp.ac.at/publications.

Finstad, K. The Usability Metric for User Experience. *Interacting with Computers*. (2010) 22 (5): 323-327 doi:10.1016/j.intcom.2010.04.004

Fostervold, K.I., Buckmann, E., Lie, I. VDU-screen filters: remedy or the ubiquitous Hawthorne effect? *International Journal of Industrial Ergonomics*, Volume 27, Issue 2, February 2001, pp. 107-118.

Gockel, B., Graf, H., Pagano, A., Pescarin, S., & Eriksson, J. (2013). An Approach to User Experience Evaluation for Virtual Museums. In *Design, User Experience, and Usability*. Design Philosophy, Methods, and Tools, Second International Conference, DUXU 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part I, pp 262-272.

Hix, D. & Gabbard, J. L. (2002). Usability Engineering of Virtual Environments. In K. M. Stanney (Ed.), *Handbook of Virtual Environments: Design, Implementation and Applications* (pp. 681-699). Mahwah, NJ: Lawrence Erlbaum Associates.

Naps, T. et al., (2003). Evaluating the educational impact of visualization. In *ACM SIGCSE Bulletin*. ACM, 124–136. Available at: http://dl.acm.org/citation.cfm?id=960540 [Accessed January 25, 2014].

Neale, H. & Nichols, S. (2001). Theme-based content analysis: a flexible method for virtual environment evaluation. *International Journal of Human-Computer Studies, 55*(2), 167-189. doi: 10.1006/ijhc.2001.0475

Nielsen, J. (1994). *Usability inspection methods*. Paper presented at the Conference companion on Human factors in computing systems, Boston, Massachusetts, United States.

Nielsen, J. & Mack, R., L. (1994). *Usability Inspection Methods*. New York: John Wiley & Sons.

Robson, C. (2002) *Real World Research*. *A Resource for Social Scientists and Practioner-Researchers. 2nd edition*. Oxford, Blackwell Publishing.

Rogers, Y., Sharp, H. & Preece, J. (2011). *Interaction design: beyond human-computer interaction*. Chichester: John Wiley and Sons

Sauro, J. & Dumas, J.S. 2009. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '09). ACM, New York, NY, USA, 1599-1608. DOI=10.1145/1518701.1518946 http://doi.acm.org/10.1145/1518701.1518946

Stufflebeam, D., (2001). Evaluation models. New directions for evaluation, 2001(89), 7–98.

Trochim, William M. *The Research Methods Knowledge Base, 2nd Edition*. Internet WWW page, at URL: <http://www.socialresearchmethods.net/kb/> (version current as of October 20, 2006 accessed 15 January 2014).