SOCIETIES
FP7-ICT
Contract no.: 257493

SOCIETIES

Deliverable D8.4

D8.4 Specification of usability testing for first prototype

Editor:	Edel Jennings
Deliverable nature:	R
Dissemination level: (Confidentiality)	PU
Contractual delivery date:	August 31, 2012.
Actual delivery date:	3rd September 2012
Suggested readers:	Usability evaluators, Trial Managers, Service Developers.
Version:	V1
Total number of pages:	52
Keywords:	User Evaluation, Field Trials, Usability, ESM, DRM, Cultural Probes, SUS, Prototypes.

Abstract

This specification for the Usability Evaluation of the SOCIETIES First Prototype includes a State of the Art on Measuring Usability for Pervasive and Social Systems, and proposed specifications for the usability evaluations for three SOCIETIES First Prototype user trials.

The State of the Art outlines the historical context for measuring usability, from human factors to participatory methods. Literature related to the challenges for usability evaluation of ubiquitous systems is discussed, with emphasis on the issues of modelling, field trials, complex organisations, and privacy. Four inquiry approaches are selected for more detailed discussion: System Usability Scale (SUS), Experimental Sampling Method (ESM), Day Reconstruction Method (DRM), and Cultural Probes.

The usability specifications for Students, Disaster Management and Enterprise First Prototype user trials are each presented, via an outline description of the proposed methodology and trial plans.

Disclaimer

Impressum

[Full project title] Self Orchestrating Community Ambient Intelligence Spaces

[Short project title] SOCIETIES

[Number and title of work-package] WP8

[Document title] D8.4 Specification of usability testing for first prototype

[Editor: Name, company] Edel Jennings, TSSG

[Work-package leader: Name, company] Mark Roddy, TSSG

[Estimation of PM spent on the Deliverable] 1PM

Copyright notice

Executive summary

This document is the specification for the Usability Evaluation of SOCIETIES First Prototype Field Trials.

The State of the Art (SOTA) in section 2 considers the historical framework for Usability as a field, and considers some challenges and distinctions particular to the usability evaluation of pervasive computing systems. We consider the extension of usability evaluation from a branch of human factors concerned with operational efficiencies, to the current focus, informed by psychology and sociology, on user experience and satisfaction. We discuss how current usability evaluation approaches are continually being challenged to keep up with increasing complexity generated by the advances in social and pervasive computing technologies.

We look at some models that have been proposed for ubiquitous usability. We describe the problems identified by earlier researchers in usability evaluations for pervasive or mobile systems that involved field trials, or complex organisations. Then we consider some privacy issues specifically pertinent to pervasive technologies. Four approaches for design inquiry are considered in detail: System Usability Scale (SUS) as a standardised survey tool; Experimental Diary Methods (ESM); Day Reconstruction Method (DRM), and Cultural Probes. In conclusions for the state-of-the-art analysis we identify the social technical model as an indication of the increasing expansion of the areas significance for usability evaluation, and refer to participatory methods as a potential for capturing future social responses.

In section 3 we outline the specification for usability evaluation of SOCIETIES first prototypes, relevant factors to be considered, and some limitations are outlined. We explain how we aim to ask users to evaluate the usability of the prototype applications for each domain, based on the project's core high-level value proposition of "Discover, Connect, and Organise", which was devised as an evocation of users' requirements and the project's research intentions. The project's innovations will be used to guide the creation of collateral for the trials.

The methodology for the trials is explained and presented, as the inclusion of SUS as a standardised usability survey that can provide a benchmark for comparison on several levels: as a temporal comparison between the first prototype and the final one; as a comparison between the applications, and user community domains; and as a indicator of usability for comparison with other pervasive systems. Secondary surveys for each group are also proposed in order to request responses and information important for the advancement of the application designs.

In the case where the trials run for longer periods, we also specify the use of a selection of secondary supporting qualitative methods, which could be diary approaches such as ESM or DRM, or perhaps more interestingly, Cultural Probes. A new variant of Cultural Probes is suggested, which extends the original material probes to include, rather than be replaced by, digital mobile probes. Specifications for each community—Disaster Management, Enterprise and Students—are then outlined.

List of authors

Company	Author
TSSG	Edel Jennings
TSSG	Mark Roddy
Intel	David McKitterick
Sintef	Jaqueline Floch
DLR	Patrick Robertson
IT-Sud	Zhiyong Yu
Heriot Watt	Elizabeth Papadopoulou
Trialog	Bruno Jean-Bart
Amitec	Yiorgos Bouloudis

Table of Contents

1 Introduction

Given what SOCIETIES has learned from previous engagements with the participants in our student, disaster management and enterprise groups to date, as described in the initial user studies documented in D2.1 and our initial user concept evaluation in D8.1, we have some contextual background to support the usability specification proposed for our first prototype user evaluation. Notably, the project has gathered a broad understanding of the participants' activities and environments, an understanding that guided the design and development of the application services and components in the first SOCIETIES system prototype. D8.2 Specification of First Prototype User Trial outlines the detailed plans for SOCIETIES first prototype trial for each user group. This deliverable, D8.4 will present the Specification of Usability Testing for First Prototype, which comprises a collection of exercises and texts to be organised in harmony with the trials described in D8.2.

In this document, we present a State of the Art on Usability Evaluation in section 2, in which a broad history of measuring usability for Information and Communication Technologies (ICT) is described; this provides a frame for outlining the shift that has occurred in usability perspectives and approaches for evaluation in the last 25 to 30 years. We consider how usability evaluation has evolved from the early strictly empirical and experimental approaches, focussing on efficiency and efficacy, to include wider human factors for evaluation, with influences from psychology and sociology. We sketch some of the challenges researchers have discovered in trying to transfer usability methods to the current "third era" of contextualized, complex, pervasive and social computing.

In section 2.1.1 we focus on examples of research done for usability evaluation for the pervasive and ubiquitous computing paradigm. We refer to work done in efforts to model usability for ubiquitous environments. We discuss issues related to bringing usability testing for mobile applications out into the field to achieve "real world" validity, and discuss the compromises and costs of doing so. Then we consider challenges for devising usability evaluation for complex organizations, looking at some issues discovered when trialling pervasive applications in a hospital setting. In addition, we focus on the connections between privacy and usability.

In section 2.2 we present some of the methods used to evaluate the users' satisfaction and experience in field trials. These include: standard usability surveys with a focus on the System Usability Scale; the Experimental Sampling Method (ESM); Day Reconstruction Method (DRM) and Cultural Probes.

In Section 3 the Usability Evaluation Methodology for SOCIETIES First Prototype is presented. Factors such as ecological validity, safeguarding privacy and trust, ethics, user motivation and access, are briefly discussed along with objectives and limits. A standardized usability survey—the SUS—is selected for inclusion in all three of the user trials; its use is expected to serve as a benchmarking tool for the usability of the SOCIETIES system. An additional secondary proprietary survey with questions specific to each community, based on research questions tabulated in D8.2, will also be distributed. Furthermore, other additional qualitative approaches such as Experimental Sampling Method (ESM) or Day Reconstruction Method (DRM) diary studies, interviews or focus groups, are to be used as appropriate for each trial, and a new variant of mixed digital and tangible Cultural Probes is proposed to open channels for alternative, open user feedback.

The Edinburgh WP8 workshop is described in section 3.7.1. During this workshop SOCIETIES researchers examined several standardized usability tests and also took part in a group investigation that involved initially matching which user activities were required to use and validate the project's third party services in development and subsequently doing a contextual walkthrough.

The details for the usability evaluation part of the three user trials are then presented for each group in sections 4 (Disaster Management First Prototype Usability Evaluation), 5 (Enterprise First Prototype Usability Evaluation), and 6 (Student First Prototype Usability Evaluation). Section 7 (Method of feedback of usability results to the project) outlines how results will be communicated within the project.

The Conclusion, in Section 8, reviews the understandings gained from the State of the Art review, the methods examined, and the methodology proposed. Finally, the Annex is used to reproduce the research questions identified in D8.2, from which the secondary questionnaires for the disaster management community, the enterprise community and the student community will be drafted.

2 State of the Art: Measuring Usability in Pervasive and Social Ecologies

This section of the document presents a State of the Art investigation into measuring usability in Pervasive and Social Ecologies.

The first section 2.1: on Measuring Usability aims to provide a broad stroke overview of the how Usability evaluation as a concept and practice has emerged from early beginnings to widespread acceptance and adoption as a key aspect of the User Centred Design (UCD) paradigm in software development processes. It traces a continuum of the usage from the scientific experimental driven methods of human factors to the more psychological and sociological qualitative approaches embraced as computing developments have extended contexts of use to complex situations.

The second section concentrates on describing some of the methods and approaches that may be relevant for usability evaluation of pervasive and social systems in the field. Approaches included are the System Usability Scale (SUS) standardised survey, Experimental Sampling Method (ESM) and Ecological Momentary Assessment (EMA); Day Reconstruction Method (DRM); and Cultural Probes.

2.1 Measuring Usability: From human factors to Field Trials

The roots of usability practice in software engineering can be traced back to the 1980s when interest in how human factors and psychology approaches could be utilised when developing computer applications grew [1]. Prior to that humans were considered as operators or programmers—to push buttons and ensure the efficient useful operation of the then large and expensive computers [1]. It has grown since then to become an established professional practice, related to User Centred Design (UCD), which is recognised to add value in the design and evaluation stages of the development of Information and Communication Technologies (ICT).

Scientific method was embedded in the lab-based engineering approach of early performance based usability tests. Software of that era tended to be designed for the workplace, with clearly defined functionality. User goals were easily defined in relation to work tasks. Interaction around the GUI assumed hierarchical importance, as a gateway to understanding Human Computer Interactions (HCI). Evaluations were developed around experiments that tested user performance, when users were asked to use an application to perform a predefined set of tasks in a lab setting. Measurements of task completion, time taken, errors, feedback and recovery were taken and variables adjusted were almost always related to the interface [2].

Subsequent to the 1980s the personal computer, World Wide Web (WWW) and mobile technologies have extended the scope of ICT applications beyond the workplace and into every facet of everyday life. Accordingly, user research studies have expanded to include social science and psychology approaches in an attempt to gain a more complete understanding of Computer Human Interactions (CHI) [1].

Usability assurance is defined in the ISO 9241-11 Human Factors standard, first drafted in 1988, in terms of 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". This definition outlines users, tasks, equipment and environment as features significant to usability, and emphasises that usability must be considered in terms of its "context of use". It indicates users goals as important targets against which the efficiency, effectiveness and satisfaction of a system should be measured [4], [5]. Context of use is specified as the scope for all usability considerations (thus encompassing the usability for complex, ubiquitous, and social contexts of use), which has continuously shifting parameters and is causing much discussion about how usability can be applied for this new current third wave of computing [6], [7].

In particular, the explosion in popularity of the World Wide Web (WWW) in the 1990s resulted in demands for new rapid approaches for usability testing to measure the value of, and help distinguish between, different designs for industry and commercial applications. Conventions about interaction modes regarding keyboards, mice and screens were established. Interactions between people and user interfaces were given great attention. New emotional dimensions for usability measurement, such as pleasure to use, introduced for consumer products, became the focus with User Experience [4], [8], [9].

In the late 1990s and early 2000s, Usability became increasingly popular, in conjunction with Information Architecture (IA), and the rise of the web design industry, was identified as a potential point for enhancing

value and offering distinction from competitors for commercial web based services. Nielson defined his techno-centric Heuristic Evaluation approach for usability experts [10], what he called a 'discount usability method,' which identified ten rules of thumb or factors for inspection by usability experts when reviewing applications. The variables considered significant for usability evaluation specified therein are:

- Visibility of system status
- Match between system and real world
- User control and freedom
- Consistency and standards
- Error prevention
- Recognition rather than recall
- Flexibility and efficiency of use
- Aesthetic and minimalist design
- Help users recognize diagnose and recover from error
- Help and documentation

At the same time, lab based usability testing methods were observed to be typically undocumented and inconsistently applied, which caused controversy over the reliability of usability methods with an experimental approach [11].

As it became more difficult to state that a particular software application was inherently and verifiably usable, and laptops and home computers became widespread, practitioners and researchers became more interested in emergent qualities of *in situ* usage to gauge usability. New toolkits of informal 'quick and dirty' practical usability evaluation methods were devised for use by an emerging profession of usability engineers in the rapidly expanding field. These included mixed methods of indirect user testing (many of which were influenced by social science observation and interpretation based research methods), such as: cognitive walkthrough [12], talk aloud [13], card sorting, ethnographic observations [14], interviews, focus groups and storyboards [15]. Krug's "Don't Make Me Think" [9] was widely read as a layperson's guide to usability. These tests allowed for some cross comparison and standardisation of usability reviews and also made some introductory level of usability tests accessible at low cost to smaller development projects.

However, as the context of use of software applications broadened beyond the well defined functionalities of an industrialised workplace, many HCI researchers indicated that to fully understand the human computer interactions it is necessary to move beyond classic usability measures of efficiency, efficacy and satisfaction, in order to give greater attention to the psychological, affective, social and physical aspects related to usability, in User Experience (UX) [3], [16]. User Experience expanded the focus of Usability to include expectations, and reflections, on experiences with technology, as well as utility and ease of use for performing specific tasks [17]. Evaluation of these cognitive aspects of usability have become more important as human computer interactions have come to be pervasive in everyday life.

The cognitive state of the human in the interaction with the computer could not be assumed to be constant, as there are huge variances between individuals in culture, prior experience, ability, interests and attention, as well as the specific situation in which the computer is used. Temporal aspects such as perceived usability and variance in usability over time for individuals was increasingly acknowledged as a factor in usability evaluation [3] [14]. What users do with the software and how they do it was measured by traditional tests, but psychologists noted it was not enough, and implied it is also important to measure outcomes—what did the user gain from the interaction—information, comprehension, entertainment, or social connection? And what did the user feel before, during and after their experience: boredom, frustration, delight, or relief? [3], [19].

Factors such as flexibility, learnability, memorability and safety were included in the early software *Quality in Use* industry standards [5], which broadened the scope for evaluating usability of software usage. The most recent update to the Quality in Use standard (ISO 25010) extends three usability factors to five qualities in use factors [2]:

- Effectiveness, measured via completion and accuracy to achieve specific goals

- Efficiency, measured by logging the time taken to do so and noting delays
- Satisfaction, broken down to sub categories: usefulness, pleasure, comfort and trust
- Freedom from risk, related to risk mitigation for economic, health and safety, and environmental aspects
- Context coverage, concerning flexibility and context range

As Social Computing is coming to the fore, a new paradigm is emerging in significance that requires further dimensions to be considered—the communal or Social Technical System (STS). In socio-technical systems, the new "user" of computing is the community. Whitworth and Friedman discuss the Social Technical System with a four level model [20]. In this model the two base levels are technical, "***hardware***" and "***software***" respectively, and the top two levels are social, namely the "***personal psychological***" level, which they associate with HCI, and the "***community***" level. This social extension of the system model is exemplified through the inclusion of norms, culture, laws, zeitgeist, group identity, customs, and myths as parts of the whole. This new extension of the system will require new usability factors for observation, measurement and validation of community experiences and usage.

Participatory methods, might be also be increasingly useful for considering usability evaluation from a social perspective, as they evolved as a democratic approach to user research in Community Supported Cooperative Work (CSCW) projects, and have been increasingly influential for user research practitioners in an effort to engage users in ongoing conversations about designs from concept creation to product delivery [21], [18], [23]. Sanders and Stappers [21] suggest that generative approaches will become more prevalent in design for complex context dependent systems, when users are the experts of their own experiences. Co-design and collective creative exercises in participatory workshops are facilitated to *"allow users to project their own needs and desires onto their imagined experiences. Artifacts, interfaces, systems and space may or may not play a supporting role in these imaginings. The ideas generated are relevant. Relevance to users means simultaneously useful, usable and desirable" [18]*

The tools for this type of co-creation in workshops tend to be pragmatic and flexible, adapted to fit particular user groups and contexts [21]. While more usually used in the discovery stage of project design, in one case group evaluation discussions, reported by Downey, uncovered seven new usability issues not elicited from the sets of individual tests, with the same participants [26].

Usability evaluation cannot be simplified to a set of methods or approaches easily reused, as each context of use is different, and each set of users are unique. It is a collection of stories, conversations, explorations and experiments involving users in consideration of the temporal trajectory of their interactions with software systems and applications from expectations to satisfaction. The diversification and extension of computing devices, uses and domains, in the socio-technical realm has left usability practitioners struggling to keep up. Some practitioners have asked if Usability is obsolete at this stage, where products are too complex, too pervasive and too easy to build [27]. In 2009, Scott wrote, "current usability work is a relic of the 1990s, an artifact of an earlier computer ecosystem, out of step with contemporary computing realities" [27]. She calls for embracing new computing tools to serve rich design methods, and employing participatory approaches. Other researchers, Buxton and Greenberg, argue that usability evaluation is sometimes not appropriate, or useful, if the right method is not applied:

"The real world is complex and ever-changing: it is likely that 'big effects' are much more appropriate for study than small ones. Yet these 'big effects' are often much more difficult to evaluate with our classic usability evaluation methods." [22]

Another practitioner Patricia Sullivan, is reported as choosing to focus on the opportunity for experimentation, and for new methods to evolve in response to emergent evaluation challenges:

"But if you're trying to develop new methods, new activities to pay attention to, if you're trying to look at usability of emerging media, then I think you still need to be experimenting. And experimenting in the sense that you're seeing each new test as a puzzle." [23]

This leads us to consider in the next section the particular issues related to usability evaluation of ubiquitous or pervasive systems.

2.1.1 Known Challenges for Usability Evaluation in Mobile, Pervasive & Complex systems

This section will consider some of the difficult challenges and activities emerging in usability investigations of context dependent, mobile or pervasive environments that are raising questions for Ux researchers. We briefly outline work done by some researchers to provide new models of usability [24], [25], [28]. Then we look at three significant challenges for the evaluation of ubiquitous computing systems that stand out from the literature, which we have chosen to discuss. The first is the need for ecological validity in pervasive and ubiquitous computing evaluations, which often leads away from lab-based testing to mobile field trials [31], [32] the second is how the essential dynamic complexity of both the ubiquitous systems and the real world complex socio-technical organisations make it difficult to draw boundaries on user evaluations [33], [34] and the third is how privacy and trust issues are pivotal in relation to usability of pervasive systems [35], [42], [36], [37], [38].

The vision of Ubiquitous Computing outlined by Weiser [39], [40] is one of inherently useable computing—not intermittently useable as ON/OFF applications, across different devices, in networks of uneven coverage, but as useable and invisible as air. Like air, such a system is effortlessly usable, unless either the system availability or the person's ability to breathe is compromised.

Weiser outlined four principles of Ubiquitous computing [39]:

- The purpose of a computer is to help you do something else.
- The best computer is a quiet, invisible servant.
- The more you can do by intuition the smarter you are; the computer should extend your *unconscious*.
- Technology should create calm.

The context of use for ubiquitous or pervasive computing is quite complex. Pervasive Computing, by definition is related to scale and complexity. It involves multiple users, tasks, devices and environments, making it difficult to measure each of the ubiquitous usability criteria for each component of a pervasive service [41]. Schotlz and Consolvo [41] developed a model for user assessment of ubiquitous applications in terms of usability and acceptance in 2004. Therein, they define seven ubiquitous computing evaluation areas (UEAs); each consisting of a series of metrics by which they could be measured:

- Attention – measured by focus and overhead;
- Adoption – which is related to the number of users required before the system becomes useable;
- Trust – understanding of privacy in relation to personal data, users understanding of the use of that data, and awareness of relationships with other users;
- Conceptual Models – this is concerned with the matching of user's mental models with the real system, which is required to support the user's ability to predict, view and understand the system behaviour,
- Interaction – this includes the efficiency, efficacy and satisfaction metrics as well as distraction, interaction transparency and collaborative interaction.
- Invisibility – this factor is measured by accountability in terms of intelligibility, controls, accuracy, appropriateness and customisation.
- Impact – related to behaviour change, social acceptance and environmental change.

While this model clearly indicates aspects of particular interest to usability practitioners, each aspect on its own would require extensive resources and time to adequately test meaning that it is not evidently easy or practical to implement in short field trials evaluating ubiquitous prototypes. Connelly [28] proposes an alternative model – the Pervasive Technology Acceptance Model (PTAM). The PTAM includes the following factors: perceived usefulness, perceived ease of use, social influence, trust, integration, and user motivation, and includes gender, age, experience and socio-economic status as moderators.

Kim et al. provide yet another model with a different set of 26 usability metrics based on ubiquitous computing, for system usability evaluation [29]. The focus of this product centred model is to measure usability performance of each component in each of the system devices to be compared, and as such support

engineering improvements. As such it has a strong engineering focus and focuses on the users satisfaction with the usage of the system in a real world setting.

2.1.1.1 Mobile Usability Studies in the Field

Field studies are by nature messy, but are often used for evaluation of ubiquitous computing projects, as it is almost impossible to model the necessary 'real world' scale of people, tasks, devices, and sensing responses in a lab setting, and often the very nature of the applications require real world contextual data to work. Researchers in mobile studies have found a need to balance between internal and external validity in user studies [30]. Internal validity, which is the focus of most lab user testing, refers to the inference of causal relationships, or how confident the observed effects can be attributed to the experimental manipulation. External validity is the validity of the generalisation of experimental findings, or how confident the observed findings can be generalised beyond the experimental setting.

In "Exiting the clean room for ubiquitous evaluation," Carter et al. [31] conducted a meta study of the approaches used by practitioners to evaluate ubiquitous computing systems in the field. Having interviewed 28 developers they found developers had difficulties developing prototypes robust enough for use in uncontrolled settings. They identified some of the key challenges as: Ambiguity and Error; Sparse data; Unobtrusiveness; and Better support for real environments, and in follow-on work proposed a new tool Momento [44] in an effort to resolve some of those challenges. Momento was a mobile tool that could simultaneously log mobile activity, take sensor readings, and cue users for feedback (using SMS and based on ESM diary methods). However, Momento could not be installed on the user's own phone, and new mobile self reporting tools emerged to leverage advances in mobile technologies that allowed event triggered cues to prompt users for responses; examples include: the Context-Aware Experience Sampling (CAES) toolkit [42] and MyExperience [45].

Antti Oulasvirta's paper on "Fielding Usability" [32] is based on the author's ten years experience of conducting field trials. It discusses in detail the limitations and challenges design and development teams faced when devising evaluation approaches for mobile and ubiquitous applications. Traditional lab based experimental methods were devised for usability to uncover causal relationships and dependencies between different system features, which assumes absolute control over all of the variables in a test environment, and the absolute randomisation of the user. However, as social and mobile technologies became more context and environment dependent, researchers began to question whether lab based experiments could ever simulate the contextual complexity required for these applications. It is difficult both to stage experiments that are truly representative, and capture the required set of data variables, using motion cameras, screen capture, and user emotional responses to the interactions.

She expands on the limitations of field trials, and cautions of unintentional researcher bias:

"Examples abound: Researchers' interactions with users may introduce biases to "natural use". The intervention creates a situation extraordinaire in which the users can use "cool" technology but at the same time are expected to do so. Also, prototypes are often isolated from relevant content and services in ways that prevent phenomena to emerge." [32]

The paper acknowledges the *imperfection* assumption necessary for field trials, while exploring the boundaries for validation and reliability. The author suggests quasi experiments as a useful approach where risks are clearly identified and principles to be tested are outlined to guide the field exercises. An example of a quasi experiment field trial where two groups of users tested the variance of usability in different interfaces of 2D and 3D mobile mapping applications in the field is provided [32].

2.1.1.2 Issues related to Usability Evaluations for Complex Organisations

Domains for proposed pervasive and social systems are in many instances those of complex organisations such as hospitals, or disaster management. There are many unknowns and unpredictable elements in evaluating usability for information and communication systems (ICT) in these situations. Users' goals are moving targets, often shifting or exploratory, and cannot be simply or easily reduced to task lists. Redish [34] distinguishes problem-solving tasks in complex organisations from well-structured tasks in the following ways:

1. Information overload is endemic;

2. Data analysis and recursive decision making is burdensome;
3. Information is often incomplete;
4. Not clear if analysis is right or wrong, in domains such as intelligence;
5. Time is critical in many domains, such as military. Getting it right can be a matter of life or death;
6. Domain experts in complex organisations may not be computer or systems experts, and they are unlikely to have much time to learn new tools;
7. Analysts and decision makers may be different people. How can data and complex systems be organised to facilitate the analysis to be visualised and presented to decision makers?
8. Visualizations are often a critical presentation method for complex information systems.

A complex organization such as a hospital raises even more complications, as Hansen reports, following experiments with embedding and using pervasive computing in a hospital [33], [43]. He identifies the following factors as significant: User Setting, Usability, Learning, Politics, Privacy, Adaptation, Trust, and Support. The following quote from Hansen gives insight into the range and types of questions complex organisations raise in relation to users of pervasive systems:

"Will end users use the system? If so, how many? Can the average user use the system? Does the interface pose problems? Does the system's overall usability match the average user? How do the users learn to use the system? Is it individual instruction or group lessons? Does the system need super-users? Is a manual or help function needed? How does the user get support? Who controls the system? Does the system change the power balance in the user setting? Who benefits from the system? Is the person that benefits from the system the same as the person that provides data to the system? Does the system require extra work from users? Does the system reveal private information? What kind of personal information does the system distribute and to whom? Is the organization ready for the system? Is there organizational resistance? Will the system change formal or informal structures in the organization? Does the user trust the system? Is the information given to users reliable? Who sends the information? Will the developers support the system? Does the support organization have remote access to the system?" [33]

Don Norman, who wrote the Disappearing Computer, which drew inspiration from Weiser's vision, recently suggested that the complexity of modern living necessitates complexity in computer systems and human computer interactions, and while good design can tame complexity, people should regard making an effort to learn or master the management of such systems, as a skill acquisition akin to reading, where learned skills are first required before people can expect to reap the rewards [46].

2.1.1.2.1 Privacy and Trust

Privacy and Trust are acknowledged as two of the most challenging and important aspects of Pervasive Computing [35], [36], [37], [38], [47]. Privacy issues are critical in relation to usability and usability evaluation, and studies show they are critical in relation to user acceptance of pervasive technologies [47].

A study by Zhu et al. [36] on identity disclosure has indicated that users are highly concerned about their personal privacy and data, but that their behaviour does not always follow their stated privacy preferences. Findings show users were not fully aware of the potential for identity disclosure via analysis of meta-data when they submitted less significant pieces of information. Privacy preferences often require lengthy selection forms, which are often ignored or hurried through by users. Based on the findings, the authors recommend that each user understands whatever interaction takes place, and that the information is understandable as well as useable. Any potential information disclosure that is feasible should be communicated to all users. Users should be aware of all third party monitoring of their data. Users require appropriate feedbacks and controls over any changes to privacy settings. Having good options to control privacy helps users feel safe.

A study undertaken by Moran and Nakata [37] to examine the factors affecting users in intelligent pervasive spaces, found that there are a set of severe recurring factors, related to Ubiquitous Monitoring (UM) identified within the monitoring and ubiquitous literatures, which are believed to have an impact on human behaviour when a system is introduced. These factors include: intrusion; awareness; control; boundaries; trust; context; and justification. They note that cultural references have sometimes caused, pervasive computing to be linked to negative connotations of surveillance, as in the cases of "big brother" and

"panoptican" concepts inspired by works from Orwell and Foucault respectively. The concept of a paternalistic system, of pervasive computing digital assistants has also been proposed as a more positive or caring vision, where horizontal non hierarchical surveillance is accommodated and even welcomed, as it supports people and helps them to self actualise.

The life logging potential and data persistence of emerging pervasive social computing also provide opportunities for *surveillance* or monitoring of the thoughts, ideas, actions, and relationships of a persons life from their self-produced autobiographical perspectives, which concerns some researchers [38]. The persistence of shared or searchable personal or social data generated and stored in social and pervasive systems presents new ethical challenges, which require ongoing investigation and research. Some researchers have argued for the right to forget and query the possibility to truly disconnect from the system to find a private space, and the tradeoffs or personal privacy sacrifices that users may have to make to access services [38].

While some of these issues may appear to be philosophical, it is significant for the user acceptance of these systems that usability is affected by expectations of use, and perceptions in broader culture, which would suggest it would be interesting for formative user evaluations to acknowledge and explore these concepts in relation to personal privacy, security and automation of pervasive systems, with users in an open ended way.

2.2 Approaches for Usability Evaluations of Pervasive Prototype

As our investigations in usability evaluation studies, in the above Section 2.1 Measuring Usability: From human factors to Field Trials show, there is a great deal of diversification in experiment and lab based performance testing versus experience based approaches for usability evaluation. There are three means of testing users experience of a prototype: experiments, inspection, and inquiry. As on-device logging is an inspection approach, which will be used for a broad range of measurements for tests detailed in D8.2, we will not discuss that approach here. The experimental or lab based approach we have learned is difficult to utilise as the full range of real world variables and conditions are not easily modelled. The quasi-experimental approach to field trials, we have also discovered is a useful compromise, particularly with regard to control groups, where mobile applications are concerned. However for the purpose of this state-of-the-art review, we are particularly interested in evaluation approaches for general usability and user experience of a pervasive and social system, rather than particular applications or products; isolating and prioritising which tasks should be included in a quasi experiment would necessitate ignoring many others. We have selected to focus on four inquiry approaches concerned with asking users to report on attitudes, experiences, concerns and visions, which will allow for a more general evaluation. Some of the following approaches are described in further detail in the follow sections: 2.2.1 Standardised Usability Questionnaires; 2.2.2 Experimental Sampling Method (ESM) and Ecological Momentary Assessment (EMA); [11] Day Reconstruction Method (DRM); and 2.2.4 Cultural Probes—Experience based HCI.

2.2.1 Standardised Usability Questionnaires

A number of questionnaires have been developed to measure the user satisfaction aspect of Usability, to support rather than replace performance testing for efficiency and effectiveness [49], [50]. Whereas the latter tests are based around the ease of use, and time taken using a range of factors, the standardised usability questionnaires have tended to use psychometrics, which have been rigorously tested and weighed with large numbers of respondents for reliability and validity [50]. Reliability refers to the repeatability of the results for a given system, with a similar selection of test subjects, and is usually calculated using Cronbach's alpha [50], which is a measure of internal reliability and can range from 0 (poor reliability) to 1 (perfect reliability). In usability practice, anything above a .70 is considered sufficiently reliable. Validity refers to the ability of the survey to actually measure and collect the data that it is designed to measure and not some other unrelated variable. Standardised surveys can be said to accurately measure the attitudes of respondents to the system tested [49].

Surveys in HCI are used to measure people's attitudes to a particular computer technology system, through querying: the user's sense of being efficient; the degree to which the user likes the system; how helpful the user feels the system is; to what extent the user feels in control of the interactions; does the user feel they can learn more about the system by using it. These surveys are of varying length, and have different response scales. Questionnaires are intended to cause the respondent to reflect on their responses, so they are also considered an intrusive approach that takes the user out of the moment or current context to respond. Some

surveys may be administered using paper and pencils, whilst others are available online. They are widely used, for a number of reasons, not least their usefulness for calculating comparable benchmarking metrics, for longitudinal user studies. Other advantages of using surveys, in addition to reliability and validity, are objectivity (reduce researcher bias), sensitivity (proved to be effective with small sample sizes), quantification (allows for statistical measurements and comparisons), economy (reuse of tool saves time and effort), communication (quantified results graph easily) and norms (some surveys can refer to universal databases enabling raw scores to be converted to percentiles) [49], [50].

As most of the well-known usability related surveys were created in the 1980s and 1990s they reflect the usability concerns most prominent at that time; in particular they are often focused on the then dominant user interface interaction features. An example is the Questionnaire for User Interface Satisfaction (QUIS) [49], which had categories for Screen, Terminology and System Information, Learning, and System capabilities. The Software Usability Measurement Inventory (SUMI) [49] is another survey of note, developed in line with the EU MUSiC project in the early 1990s, which combined aspects from pre-existing surveys, like the System Usability Scale (SUS) [51] and QUIS, with factors arising from discussions with developers and end users. The WAMMI [49] is designed for evaluation of web-based applications. The Mobile Phone Usability Questionnaire MPUQ [52] is a questionnaire developed to gather subjective usability information about mobile devices, which has 72 questions grouped into six different categories.

The System Usability Scale (SUS) was one of the first usability questionnaires, and its widespread use and broad questions have allowed the validity and reach of its results to be extended beyond the original intention [51], [53] and it has been found by Tullis and Stetson [54] to psychometrically outperform other subjective scales including QUIS [54].

2.2.1.1 System Usability Scale (SUS)

The System Usability Scale (SUS) is a usability satisfaction questionnaire, which was developed by John Brooke, while he was working for Digital Equipment Corporation in 1986, as a ‘quick and dirty’ attitude evaluation method, for electronic office systems [51], [53]. It has been made freely available for use in usability assessment. It is a short questionnaire and quick to complete, which meant it could be applied in demanding industry contexts where test subjects had limited time. Since then it has been widely adopted in a range of projects from software applications, websites, to yellow pages and mobile applications, and has become accepted as an industry standard with references in over 600 publications [53].

SUS consists of the following ten statements, five of which are positive and five of which are negative. Each user is invited to evaluate the statements in relation to the system in question, using a Likert scale. The statements are:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system

Users respond to the statements by ticking a box on a Likert scale from 1 to 5 to indicate whether they “strongly agree” (1) or “strongly disagree” (5).

The SUS scale is generally used after the respondent has had an opportunity to use the system under evaluation, but before any debriefing or discussion takes place. Respondents should be asked to record their immediate response to each item, rather than thinking about items for a long time.

SUS is scored quantitatively according to the following rules.

1. For odd items: subtract one from the user response.
2. For even-numbered items: subtract the user responses from 5.
3. This scales all values from 0 to 4 (with four being the most positive response).

Add up the converted responses for each user and multiply that total by 2.5. This converts the range of possible values from 0 to 100 instead of from 0 to 40.

An extensive empirical evaluation of the System Usability Scale was conducted by Bangor et al. [45], who concluded that it is "a highly robust and versatile tool" which "could quickly and easily collect a user's subjective rating of a product's usability".

Sauro makes the following observations having reviewed and analysed data from over 5000 users across 500 different SUS evaluations [53]. He deduces that SUS is a reliable and valid measure of perceived usability, which can detect differences at smaller sample sizes as well or better than commercial questionnaires and proprietary questionnaires. He recommends the use of mathematical formulae that calculate the mean and standard derivation, when sample sizes are small, in any given study. The standard derivation is the size of the variance from the actual results, and the confidence interval gauges how realistically indicative the test sample results are. In addition, recent research indicates that while it was only designed to give general usability indications, SUS has been shown to also provide a global measure of system satisfaction and sub-scales of usability and learnability. Items 4 and 10 provide the learnability dimension; and the other 8 items provide the usability dimension [54]. SUS is not however diagnostic and will not illuminate the cause of usability problems. It is useful as a benchmarking tool to facilitate cross comparison of results from different systems or stages in the development cycle, making it ideal as a benchmarking tool for ***first and second prototypes***. Analysis and presentation of SUS results is further simplified though the provision of calculators for a small fee [53].

While summative tests are often based on quantitative tests and qualitative tests for formative testing, even small numbers of users can provide relevant quantitative results for task completion and task time at formative testing stages, if confidence intervals and binomial distribution are used with the Adjusted-Wald Binomial Confidence Interval [55]. These are mathematical formulae that calculate the mean standard deviation, based on the sample size, in any given study, meaning realistically indicative test results can be calculated from small samples and confidence intervals.

2.2.2 Experimental Sampling Method (ESM) and Ecological Momentary Assessment (EMA)

Experimental Sampling Method (ESM) and Ecological Momentary Assessment (EMA) are methods devised in cognitive and behavioural science to intermittently capture *in situ* phenomena from people in their natural environments, which have been adapted by HCI for evaluation, and are often casually interpreted or referred to as mobile diaries, or life logging. ESM was an experimental method developed by Mihaly Csikszentmihalyi in the 1970's [56]. It aims to capture empirically self-reported data as it occurs, thus reducing the cognitive bias that occurs with reporting from memory or reflection. It depends on substantial repeat observations, the timing of which is significant to each study.

Users are contacted or sampled by researchers at regular, random or contextually decided intervals, and asked to report on their thoughts. At the signal the person writes down where she is, what she is doing, what she is thinking about, who she is with, and then rates her conscious mental and emotional state at the moment on various numerical scales: how happy she is, how much she is concentrating, how strongly she is motivated, how high her self-esteem is, and so on. The responses are episodic or experimental representations of the user's experiences at a particular point in time rather than semantic or generalized. Data derived from ESM studies is empirically analysed with variants in of frequency and patterning of activities, social interactions, location, psychological states or moods, and thoughts. Originally ESM requested that respondents use a page or watch with a diary or notebook to record their experiences but the tools for self reporting have naturally advanced as new technologies emerged.

ESM has been adopted for user studies particularly for ubiquitous computing applications, as a method for learning about users 'in the wild' [25], [42], [44], [45]. It is often used in conjunction with logging, as it is said that logging will provide data on the what, where and when but ESM is required to learn about the whys of a given situation. ESM is concerned with subjective data about user perception, intention and satisfaction. Several specific tools have been developed to support the method notably CAES [42], Momento [44] and My

Experience [45]. Context-Aware Experience Sampling (CAES) was devised by MIT researchers in 2003, to take advantage of the affordances offered by sensor technology, to cue requests for user feedback based on user interactions with the system under investigation. So for example a user will be asked for a response just after the use of particular functionality or service [42].

An advantage of ESM is that it captures a participants experience in the moment and environment of doing, thus reducing the biases of recall. Of course, responses are consciously filtered and participants can choose what to reveal or conceal. It is dependent on participants agreeing to be actively involved and motivated to share, and is therefore considered to be quite an intrusive method. Sensitivity to the limitations of users to respond in particular circumstances or environments should be taken into consideration. As with all studies that seek to understand users outside of formally coded environments such as work, there is a negotiation between participants personal space and privacy and the project study, which need to be carefully negotiated [25].

2.2.3 Day Reconstruction Method (DRM)

The Day Reconstruction Method (DRM) is a diary method sourced from cognitive science, which combines aspects of Time-Budget Measurement and Experience Sampling Methods [57]. DRM seeks to measure reflected experiences. It is not as popular as ESM but has been used in mobile studies to measure user experience over time. Martens et al. argue that these post experience reflections are as significant if not more so, than in situ expressions, as they reveal the formed opinions, impression and biases, which influence future decisions and interactions with the system [58].

In this method [57], users are briefed during an introductory session on the method, and tools for its application. They are asked to keep a private notebook in which they record their personal recollections and opinions about their use of the system, and to which they may refer when completing another reporting exercise the following day, or in some cases at the end of the day. The private notebook is for the individual participants use only and is not accessed or collected by the researchers. Its purpose is as an *aide memoire*. A second notebook is then given to the participants for their reporting exercise. The reporting exercise is to be completed after the personal recollections, and the users should not know the specific nature of these reporting questions before they complete their notebooks, as they may bias their observations. The reporting exercise requests each user to recollect the three most memorable interactions, regarding the system being evaluated. These questions ask them to describe key features of each episode, including (1) when the episode began and ended, (2) what they were doing, (3) where they were, (4) whom they were interacting with, and (5) how they felt on multiple affect dimensions. The response form is returned to the researcher for analysis. The DRM can be conducted in an individual or group setting, and is reported to take between 45 to 70 minutes to complete. Results are interpreted after traditional content analysis is preformed [57].

2.2.4 Cultural Probes—Experience based HCI

"*Cultural Probes are designed objects, physical packets containing open-ended, provocative and oblique tasks to support early participant engagement with the design process."* [59]

Cultural Probes is a design inquiry approach, inspired by the Situationalist art movement and devised by Gaver in 1992, through work done for the EU funded Presence project [60] [59]. The probes emerged to answer the design team's desire to move beyond stereotypes and develop sensitivity in listening to other people and understanding their actions. The probes were designed to invoke user agency by provoking reflective responses, and designer agency through their thoughtful creation and interpretation. The goal of the Cultural Probes method is to elicit inspirational responses from people rather than collect comprehensive information about the respondents [59]. These responses are cultural artefacts, which can inspire designers, with clues about the respondents' thoughts, actions and environments. The method's focus is on empathic and participatory design, embracing the idea of informing design through open and playful discovery through creative exchanges [59].

Probes are distinctive from other qualitative research approaches in HCI. The probes intent differs from ethnographic methods in the sense that it does not aim to be a non-intrusive observational data collection and interpretation method to inform cultural engagement; rather it is a process to involve users in a design conversation by making things. It is similar to ethnographic methods in that both approaches focus on the

significance of the user as an expert in communicating and disclosing cultural and contextual information. This dialogic conversation between designers and users employing Cultural Probes is similar to Participatory Design, but in Cultural Probes design team retain power through setting constraints in the design of what is to be included in the probe kits, and also as sole interpreters of the significance of the returned objects [61].

The original tangible Cultural Probe kits are made up of material objects, presented as a packet of evocative tasks and items, to the people who agree to participate in the project. The package contents are experimental in nature and uniquely tailored to each project. These packages often include artefacts to support exploratory tasks and could, for example, be composed of maps, diaries, disposable cameras, stickers and envelopes. They could also contain images, stories or objects to stimulate the respondents to reflect on their situation. In a domestic probes project disposable cameras were distributed, and participants were asked to photograph "something you'd like to get rid of," "the spiritual centre of your home," and "something red." [62]. Design teams are encouraged to include open-ended questions, which would spark a user's curiosity and imagination. Respondents may choose to engage, or not. Each respondent is free to express his or her responses to the Cultural Probes in her or his own unique way. People responding to probes may engage in lots of different activities like, gathering or making images, making notations on maps, writing on postcards, keeping diaries.

The collection of probe responses is later sent back to the design team, sometimes in trickles. The responses may take the forms of images, stories, diaries, audio clips or video, and are often imaginative and exploratory in nature, forming a rich set of cultural indicators, which the design team sifts through for inspiration, without further explanation or clarification from the respondents. Scientific tools or approaches for data analysis are not used. Gaver emphasises the value of this approach is in the ambiguity of the responses [62]. The unexpected might never be discovered if the probes are overly directed and constrained [61].

"More definite sources of information seem to close down possibilities, to provide rigid parameters for design. … Probes create a productive balance between confronting us with lots of things we didn't know about or expect, while leaving huge amounts of room for us to move."[63]

HCI has embraced the Cultural Probes approach as a useful tool for glimpsing moments of understanding of user environments, and experiences, as technologies have become interwoven into the fabric of everyday life, and researchers seek new methods to gain access to users private worlds [59]. The probes approach has developed so many variants with use that it is better described as a collection of methods. However, in some cases where projects attempt to employ Cultural Probes in a formulaic or recipe-like manner, the subversive and reflective intention of the original method is reduced to a means of data collection. Boehner et al. de-construct how the tendency of practitioners to veer away from the ambiguous multiplicities of the original method in their implementations, indicates an underlying tension between different types of knowledge making, in the field of HCI, dominated by use of epistemological approaches [59].

2.2.4.1 Digital Cultural Probes and Mobile Probes

Some interesting variants of the probes are those distinguished by the media used for their dissemination. As smart phone applications have been embraced by users for a myriad of tasks, the phones themselves are often ordinarily used to record, capture, log people's ideas, thoughts and experiences, in diaries, lists, photographs, videos and social interactions. Individuals' close relationships with their mobile phones is what has inspired such interest in the potential of using phones for participatory, opportunistic, or human-centric sensing and it is this same aspect of their use that has lead to the suggestion that they are potentially ideal tools for probing. When utilised for user studies with mobiles, design probes have sometimes become characterised as self reporting tools, that allow users to give 'in the moment' access and responses to researchers tasks and questions, as described by researchers from the University of Art and Design Helsinki, who employed probes in several projects [64], [65].

Mobile Probes is a variant that aims to simplify the probes approach by digitising all the data [65]. In a case study from 2004 mobile probes were used as a tool for facilitating a study into shopping culture, with shoppers, and store assistants The exploratory nature of the probes was used in conjunction with a more formal ESM tool, where users were contacted by SMS at regular intervals to answer predefined questions. The digital responses were collected on a website and then later tagged according to time, user and question, printed and used in participatory workshops to create collages individually, and elicit more general findings in group discussions. Researchers noted that the use of the digital web database platform as a gathering zone

for responses supported automatic categorisation, reduced researcher anxiety about the level of responses, centralised comments on the responses, facilitated organising data for workshop exercises, allowed all participants ongoing access to the material, and provided powerful conversation drivers in the participatory workshops [65]. These distinctions from tangible probes are seen as advantages.

Digital Cultural Probes is another media defined variant, which was used in the BRIDGE project as a means to involve children with participatory methods [64]. The children were given mobile phones, and invited to take photographs, make audio clips and send messages, which were then shared in a closed online forum. The children's informal and spontaneous responses gave the researchers insight to the children's practice and learning, and provided them with a space for exchanging inspirational ideas.

2.3 Summary of State-of-the-Art Review

We considered how the emergence of usability as a practice has been influenced and altered by many different disciplines from engineering science, psychology, sociology and the commercial context of its application [1]. We discussed how usability is defined (in ISO standards) and varies according to usefulness and usage, in specific contexts by specific people with specific goals [5]. Human aims, thoughts, attitudes, impressions and intentions, when interacting with computer systems have become of increasing interest to interaction designers and developers seeking to understand deeper human qualities and desires in assessing the value of new computer systems. Expectations and reflections of use have been recognised as significant user experience elements not easily separated from usability, when the actual usage of a system is required. Social computing introduces more factors for consideration in systems design and evaluation, as culture, norms, relationships, group identities, can all influence acceptance, adoption, maintenance and spread. Participatory methods are looked to as an inclusive and expansive approach, where creative co-design approaches may be adapted for formative evaluation purposes. Usability does not have a definite clear set of methods and models for the evaluation of pervasive and social computing applications or systems, which are easily applied.

However, having made reference to some of the more traditional models and standards for usability, we then looked to some the challenges specific to ubiquitous computing. Some of the new models proposed for usability evaluation in pervasive or ubiquitous computing were outlined.

We then focused on literature related to three aspect to usability evaluation, which we felt were particularly interesting, as the widespread popularity of mobile social and pervasive applications has redrawn the borders of usage, from clearly defined tasks to unpredictable activities in everyday life, for ubiquitous computing systems. Firstly we discussed evaluations for mobile applications in field trials, then usability for ubiquitous computing in complex organisations, and finally the significance of Privacy and Trust for user acceptance in all ubiquitous computing usability considerations.

In reference to the literature on field trials, we learned that while field studies are increasingly necessary for mobile and pervasive application, they reduce the experimenters ability to control variables in the environment [32], and that this leads us to consider the value of different approaches to knowledge making in usability studies [31]. An account of implementation of a ubiquitous system in a hospital exemplified the multiple layers complexity of users in the system, and the importance of expert users input regarding work practises [33]. Privacy and trust issues, while hard to pin down, were briefly discussed, as they are acknowledged as pivotal to usability and user acceptance of pervasive and ubiquitous systems [36]. We also considered research indicating that ubiquitous monitoring in pervasive environments may itself cause users to alter their behaviours [37], and broader concerns raised in relation to technologies gathering of personal information [35], [38].

The second part of the SOTA investigated a selection of four tools and approaches that could be used to evaluate usability for mobile social and pervasive systems in field trials: SUS, ESM, DRM and Cultural Probes [51], [56], [57], [60]. The SUS was given particular focus as a verifiable and well documented and trialled usability evaluation survey, which can give a generalised usability metric, and thus is useful in terms of benchmarking, and reliability for longitudinal studies [51], [53]. An additional advantage being that it is short and easily facilitated in a test situation. ESM [56] and DRM [57], [58] methods were described as useful approaches for in-depth user studies of technology in use. However, it is acknowledged that such diary approaches are intrusive and demanding for users to complete, and therefore have a dependence of selecting very motivated and diligent users for participation.

Cultural Probes [25], [56] were discussed in detail, as a design inquiry approach to provide space for ambiguity through multiplicity of stories and perspectives in the design space [57]. The discussion referred to two variants in particular: Digital Cultural Probes [57], and Mobile Probes [60], while referring to the literature, which uses the use of probes in HCI to illustrate and discuss both empirical and rhetorical approaches in the field.

The next section of this document, Section 3 [67], will outline the specification for the usability evaluation of SOCIETIES First Prototypes for the disaster management community, the enterprise community and the student community.

3 Usability Evaluation for SOCIETIES first prototype

"It is particularly important to ensure a tight coupling between the claims you make about your system and the evaluation methods that you use to demonstrate that these claims hold."[66]

Prototyping ubiquitous systems is a well defined development stage, recommended by Weiser. Allowing users access to the system via third party service applications, is a good way to test system features, as long as those applications embody careful design and can be shown to rely on system components [66]. Similarly it would be easy to evaluate the prototype applications for usability, and not focus on the usability of the entire system. A negative usability result for an application does not necessarily mean the system is unusable. Similarly a false positive could occur. Therefore with SOCIETIES prototypes, we should also test and measure the entire system against the project innovations and value propositions, not just individual application successes. How easy is it to discover, connect and organise with other relevant people, services and entities across the digital and real worlds which SOCIETIES spans? How does it feel and what do users explicitly and tacitly think about belonging or being excluded from a Pervasive Community? What are their experiences of our twelve proposed technical project innovations: Personalisation, Learning, Community Preferences, User Intent, Relevance, Community Orchestration, Context, Community Context, Location, Safeguarding, Privacy and Trust?

As the SOCIETIES applications and supporting pervasive platform functionalities are still in development, these technologies are not mature and the interfaces are not polished, users will not at this stage experience the fully immersive pervasive environment where they can discover, connect and organise with relevant people and entities via pervasive communities, which is the vision towards which the SOCIETIES project is striving. Therefore the usability evaluation of these immature prototypes is formative, with the goal of highlighting known and unknown errors, limitations and concerns with the emerging system and technologies from the users' perspectives. User engagement at this formative stage, should be employed to inform future design decisions and development directions from the users' point of view. This usability evaluation stage is nevertheless necessary to promote and engage with the human perspective, and consult with the people who have helped shape the project vision, inspire and define the applications, and ensure that human smartness [68] remains to the fore in ongoing project considerations.

This evaluation stage also raises many questions. Can the SOCIETIES system deliver on the promised values of Discovery, Connection and Organisation to these early users, if the proposed project innovations do not have the required minimum data sets about the users to work or to work well? If not, will the users think the applications designed to leverage the platform features nevertheless meet their needs and requirements?

How might satisfaction, freedom from risk and context coverage be evaluated then if we allow for more factors of Quality in Use (ISO 25010) standards be acknowledged in SOCIETIES first prototype trials? The qualitative diary methods from *in situ* user reported ESM entries, or DRM recollected interactions might be applicable, if users who are sufficiently motivated to engage in that level of activity can be found within the user groups. However, as we will not have access to negotiate this with any users prior to the trials, it is risky at this stage, to assume several lead informants will emerge from each of the groups. The large amounts of data these approaches generate would also create a demanding analysis task.

Additionally, as much as the objective of the SOCIETIES project is to enable people to discover, connect and organise across virtual and real worlds, and ubiquitous computing aims to pre-empt users' requirements and wishes, any attempt to evaluate the usability of the projects applications, should acknowledge that the uniqueness of personalized experiences in a particular combination of contextual settings, makes generalisability and repeatability virtually impossible.

3.1 Scope for Specification of Usability Evaluation for First Prototypes

Measuring usability is necessary in order to both communicate where future design effort is required and also to provide benchmarks to indicate how future alterations and improvements post the trials improve usability between rounds of iterative usability evaluations. Whilst, in product design it would be optimal to aim for several rounds of iterative testing, in the SOCIETIES project we are restricted to the predefined availability of our user groups, within the three rounds of user trials and evaluations that have been planned. The first round of user evaluations as described in D8.1, comprised of the qualitative storyboard evaluations which tested the project concepts in focussed discussions with representative groups of users from each of

our three user groups, Students, Disaster Management, and Enterprise. This next round of user evaluations will aim to capture both user expectations and responses to the first SOCIETIES prototype, including those tests documented in the D8.2. The purpose of this document is to specify only the subset of usability evaluations of these prototype trials. The document D8.2 provides an in-depth description of the trials, and is designated to be a living document, which is updated as details become cemented closer to the trials, so the object and scope of this document is to complement D8.2, not to replace it.

As the time with users is limited, and the services and system are incomplete, the principal value for which we can hope to achieve, from these first prototype usability evaluations will be to gauge the users' general perceptions of the prototyped applications and Pervasive Communities system features, while also highlighting any major design oversights and flaws.

It is beyond the scope of this task at this stage in SOCIETIES development to specify all the feasible usability task tests for each of the third party services or subsystem components. The logging specifications pertinent to each particular set of services, for each of the three user domains will be detailed in D8.2 updates, and these activity records will also provide data relevant to usability evaluation of efficiency and efficacy, in terms of successful key task completion, time taken to complete tasks, calls for support, mutual adaptation of system and user components with personalisation and learning, errors and recovery from errors.

3.2 Ecological Validity and Field Trials

The notion of ecological validity for user trials, as discussed in the SOTA, is pertinent for SOCIETIES as with most mobile and pervasive systems. This concept relates to user testing being conducted in a real envionment, where users are immersed in the surroundings and distractions of their everyday normal lives, as opposed to taking on the role of a tester in a laboratory setting. The objective of aiming to have ecologically valid tests is to capture users' interactions with the system, amidst the full range of complex interpersonal interactions, and environmental factors that provide rich context information to the humans in situ, affecting their experiences, as well as subsets of complex contextual data relevant for the pervasive communities services. Laboratory tests are easier to control but the results while they might create nice graphs, are not so easily mapped to real life, where unknown unpredicatbles can impact on the user experience.

The Student and Enterprise trials will take place in a natural environment of each communitity, a university and a conference. The Disaster Management trial will take place with a simulation of a disaster event. This highlights a major difference between the Student and Enterprise trials on the one hand, and the Disaster Management Trial on the other: it is impossible (if not undesirable) to recreate the full extent of the natural environment, in particular at the user level. Whilst DM experts may be accustomed to emulating DM scenarios and their burdens through training, this may be much more difficult for the volunteer group.

For the puposes of the SOCIETIES trials, users should be made aware of the limitations of the prototypes, and the distinction between a prototype and a fully working system. Full ecological validity will not be possible in the case where parts of the conceptualised system in the project scenarios, on which the first prototypes are based are not working. The interconnectivity of all of the disparate parts is essential to a pervasive system, which means that there is a risk that a lack of integration could significantly impede usability measured at this stage. It is expected however that SOCIETIES value proposition for each domain, will be evident from the services prototypes, and that the users will be able to evaluate if the proptotyped services fit their needs, and suggest improvements.

As in all field trials, the potential for evaluation measurement tools and the process of participation in a trial to have an effect on the "natural" behaviour of the participants, may be expected to introduce some aspects of bias [69], and a balance will have to be carefully negotiated to facilitate the varying research interests in SOCIETIES in the trials, to both monitor natural behaviour within the system, and capture participants reflections on their experiences.

3.3 User Motivation, Engagement and Ethics

The project acknowledges that engaging users to participate may not be easy, as users of all groups have busy schedules and prior commitments. Similarly motivating people to use the services might be difficult, if

the service recommendations do not have a minimum required data available to make truly useful suggestions or decisions.

The *a priori* identification of SOCIETIES researchers with other organizations from the users' perspectives is both an advantage in facilitating trusted access which might not otherwise be possible, and a disadvantage in that the social obligations and power relations such relationships entail might potentially introduce an unintentional bias from the researchers or participants. Alternative channels for engaging with users in a participatory manner, such as Cultural Probes, could potentially facilitate anonymity and provoke users to engage with the more abstract project concepts, and share tacit cultural information.

The sampling of people who will participate from each of the three user groups will be self selected, or by convenience selection, and thus may not be entirely representative.

The ethics of the trial will be dealt with, as were all previous project engagements with end users. (C.f. reference to ethics in D8.2, D8.1 and D2.1) Users will be informed about all monitoring and data management policies of the project. Information regarding data collection, storage and deletion will be provided before they begin the trial. Users are free to cease participation with the trial at any point.

3.4 Privacy and Trust

Some of the HCI literature, in relation to pervasive computing, we have noted, has observed a tension between usability and privacy [35], [36], [37], [38], [47]. Privacy and Trust are mentioned as key non-functional aspects of user acceptance of Pervasive computing [66]. This quandary has emerged in SOCIETIES user research and storyboard discussions: How could the users know if the system is trustworthy if they have not yet experienced it as it does not yet exist? In this case, would they not be well advised not to disclose personal information before they have an opportunity to explore and learn more about it?

Managing communications with potential users about privacy protection, security, data management, data persistence, risk warnings and flows of information within the prototype system is likely to be significant in the formation of users' attitudes to the system, before, during and after the trial. Trial participants need to know what personal information will be collected, and stored, where and for how long. Participants need to understand which aspects of the system are available for interaction. They also need to be updated about context changes, and have visibility of their status and related data flows within the system.

SOCIETIES will provide understandable and informative consent forms to all users at the beginning of the trials.

3.5 Temporal Factors

The differences between the experiences that people have using any computer system or applications have a temporal variable, referred to in the literature as the experience trajectory [68]. First time users of a service are recognized to have different needs to discover, learn about and explore the features offered. Designers often recommend using progressive disclosure for large complex systems, so that the interfaces with which user interact, are responsive to the users' level of knowledge and familiarity with a given system. Progressive disclosure reduces cognitive load on users. However in SOCIETIES, even if the trial participants prefer to share data generously, and are eager to explore the prototype services, the benefits of Pervasive Communities would be progressively experienced. Pervasive systems are evolutionary. The pervasive experience could be expected to be a little bumpy initially, as it monitored and learned the preferences of individuals and communities, but the vision of users and communities having delightful experiences, plain sailing with invisible technologies supporting one and facilitating relevant connections with people and services continuously throughout the day, is unlikely to be feasible with the necessary limitations of a first prototype, with unripe and undigested data. So the user trials are challenged to both manage the expectations of users regarding the reality of the maturity of the Pervasive Communities system, all the while convincing them that their participation will be worthwhile, interesting, and rewarding if they choose to share data and enable monitoring, although SOCIETIES will not be available long enough for this to be possible. This expectation that users' experiences of Pervasive Communities would be enhanced over time, as they key innovative and distinctive features are uncovered and empowered, would indicate that the shorter trials, could not realistically predict positive usability results, that might otherwise become measurable over time. The longer Student Trial is of particular interest for this reason.

3.6 Overview of Trials and Services

The first prototype usability evaluations are formative[1] in nature, and they will be field trials situated in the *milieu* of a university, a disaster management trial and a technology conference. The first prototypes to be tested in the trials, will be accessible to users within application specific use cases, which are based on services particular to the disaster management, student and enterprise communities.

SOCIETIES first prototypes will be presented to volunteer participants in each of the three user groups for the duration of each groups' User Trials, as described in detail in D8.2. Three distinctive sets of third party service applications have been developed within the project to meet the requirements of each group. The duration, location, environment, and culture of the three groups participating are all variables that vary greatly for these trials. These factors combined mean that three different specifications are required to manage usability tests for each group.

The disaster management trial will be a simulation of a disaster assessment situation, where a quasi experiment will be set up over two phases. The participants will be invited/instructed to utilise the i-disaster, AnalyzeThis, IWantToHelp, YouRNotAlone and Disaster Data Collector services.

The student trial is planned to take place in Heriot Watt University in Edinburgh in early 2013. The participants will be selected from the same student group that has already informed the two initial user engagement activities in SOCIETIES. They will be given smart phones and invited to utilise the SOCIETIES services, which have been proposed by the project as to leverage the features and innovations of the Pervasive Communities platform in their university milieu. These students' interactions with the applications and devices will be captured by logging, and other experimental laboratory tests outlined in D8.2. In addition this document specifies tests that are planned to test the current status of the first prototype applications' usability. The SOCIETIES third party services which will be offered are: CoBrowse, EventHerald, CrowdTasking, NearMe, SocialLearningGame, MyTV / Pesonalised Display.

The project plans to negotiate with these participants to find out if they are willing to be involved in additional qualitative research or design inquiry activities, such as Cultural Probes or Day Reconstruction Method, and if so these additional activities may also be undertaken to inform and improve the usability of the applications for the final prototype.

The student first prototype trial is distinctive from the other SOCIETIES user trials: firstly, it is planned to take place over a much longer duration, a period of weeks. This extended duration of the trial will definitely mean that the organization, facilitation and support required for the trial will be more extensive and demanding than for the other trials, and the risks for failures on several levels, technical, social and personal are increased. Students will have access to the prototype SOCIETIES platform APIs and tutorials as well the tailored student applications, but the demands of their course work may inhibit their availability to experiment extensively with them. It also introduces further work and risk, as supporting documentation, guidelines for use, and live support for developers will be required.

The Enterprise trial will take place in February 2013. The exact date is not yet decided. It will be a field trial that will take place at a technical workshop or conference. Delegates will be invited to trial the SOCIETIES conference services: AdHocMeetings/CollaborationTools, NetworkingZones, SharedCalendar / PersonalizedAgenda, Context Aware Wall, and Conference Registration.

3.7 Selection Process for the Methodology

In an effort to introduce a benchmarking tool, as well as indicate a global evaluation across the three groups, which can be validated in relation to external data, it has been decided to include a standardised usability test. However having reviewed a number of the standard usability tests available, it is clear that most were designed with the expectation that they would be used in the context of gathering usability data about

[1] Formative usability approaches are those activities that take place during the iterative design stages of the project development cycle, at points when the design prototypes require feedback to guide development decisions and help plan next steps.

applications served to web, desktop, or mobile interfaces, and these would not therefore be suited to evaluating the disappearing or widely dispersed flux of interfaces across multiples of devices and entities envisioned by SOCIETIES Pervasive Communities applications, without extensive adjustment. However even minor adjustments have been calculated to affect the validity of standardised tests [51], [53]. Whilst the shift to mobile communications and smart phone applications has led to new usability surveys being developed for remote and mobile survey testing [51], [53], the content of these survey questions are often based on interface variables. However GUI controls are not central to SOCIETIES core value proposition of Discover, Connect and Organise.

We organised a workshop to explore the questionnaires, as described in the following section, and it was decided to use the System Usability Scale (SUS) for SOCIETIES First Prototypes Usability Testing.

However, as the SUS, while valid and reliable, is not tailored to SOCIETIES specific concerns, and we also have already generated questions specific to each community regarding the use and experience of the prototype services, therefore we also propose to have an additional questionnaire for each community. These secondary questionnaires will be drawn, mostly from the questions identified in D8.2 and reproduced in the annex of this document. The length of this questionnaire will be limited for practical purposes and to avoid overburdening the users.

It is expected that the results to these more specific questions will inform and guide development for further enhancement of SOCIETIES' prototypes. However due to the small sampling and difficulties associated with establishing validity with the survey method, it is acknowledged that these results may not be necessarily repeatable, reliable or generalisable outside of the specific confines of the SOCIETIES project.

3.7.1 Edinburgh Workshop

At a SOCIETIES plenary meeting in Edinburgh in June 2012, a WP8 workshop was organised to discuss potential surveys for measuring usability at the First Prototype User Trials, and to discover researchers' assumptions about user activities.

Researchers were asked to attempt to respond to sample questionnaires based on a previous media product of their personal choosing. Four different standard questionnaires were considered:

1. Mobile Phone Usability Questionnaire (MPUQ). [52]
2. System Usability Scale (SUS). [51]
3. Software Usability Measurement Inventory (SUMI). [70]
4. Website Analysis and Measurement Inventory (WAMMI). [67]

Each of the tests had been shown to produce valid quantitative results, and could therefore offer some benchmarking results. Project researchers present at the workshop noted they observed several severe limitations to the adaptability of three of the tests to measure usability factors for the SOCIETIES first prototype. The WAMMI was felt to be more suited to website evaluations, with many questions irrelevant to the expected experience of a SOCIETIES' user [67]. SUMI is a psychometric test that was developed to test end users quality of use for software [70]. It has been rigorously tested and proven method of measuring software quality from the end user's point of view. However SUMI was thought to be too long, at 55 questions, for practical use in SOCIETIES trials. The MPUQ was considered too lengthy and specific to device affordances for SOCIETIES requirements [52].

The workshop approach gave researchers an insight into the experience of trying to answer non-specific questions, and introduced them to the types of questions available with standardised and validated questionnaires, with which the usability of the SOCIETIES system would be measured.

It was decided, during this workshop, to use the System Usability Scale (SUS) [70].

This is one of the shortest standardised surveys so it will be quick and easy to complete, which is as much a factor for SOCIETIES usability tests for the first prototype as it was for Brooke and his industry test participants and clients. The questions are quite general and do not overly emphasise feed in relation to interface issues. We can ascertain that it is a reliable test, having been proven to be so after its widespread use and statistical investigation of results from thousands of SUS surveys. It is also free, and well-

documented support, for administering the test and clear guidelines and tools for analysing the data collected, is available. The results for a set of user responses can be calculated as a single global score, which means that the SUS, is ideal to use also as a benchmarking tool, for comparative usability across the results for the three user groups in SOCIETIES, and also as an easily repeated test, provides a useful measure for longitudinal comparisons within groups. Studies have shown SUS results prove a strong correlation between Usability and Learnability [54]

However the SUS is limited as the questions are very generic, and the results it calculates are unlikely to generate specific usability recommendations. Therefore, a secondary questionnaire composed of a short set of questions for each community, will be drafted from the qualitative research questions formulated in D8.2, and reproduced in the annex of this document. (See Annex A: A.1 Formal Performance Metrics and Feedback for WP4, 5 & 6, A.2 Qualitative User Feedback from Students, and A.3 Qualitative User Feedback from Enterprise Community).

Reviewing the surveys raised some questions regarding Usability boundaries: which aspects of SOCIETIES 3P services for example would be understood by users as relevant for answering the questions? This discussion raised some interesting challenges with regard to how variances within SOCIETIES services and the platform could be evaluated. Will the results from our usability tests at the user trials, be granular enough, to recommend which aspects of the prototype platform require usability improvements?

A second exercise was organised as part of this workshop, around expected user activities. Developers, necessarily make some assumptions about user motivations, behaviours and desires when creating services. While the SOCIETIES project has made considerable effort to involve, and include the people for whom we are developing services in the three domains of Disaster Management, Students, and Enterprise, as we have documented in D2.1, D8.1 and D8.2; there are still considerable gaps in the project developers' knowledge and understanding about the habits, preferences, routines, and goals of people in the project's target user groups.

During the prototype trials it is expected that users will form their opinions about the first prototype based on their interactions and experiences with the third party services, which will leverage the innovations and features of the SOCIETIES platform. Therefore, it made sense to collectively consider, how the services fitted or suited the lives and requirements of the people participating in the trials.

The attendees of the workshop split into three groups, with the objective of listing the types of user activities and tasks required, to set up, use and validate each of SOCIETIES' third party services. One group focused on Student Pervasive Communities services and student activities. Another focused on Enterprise services and activities, and yet another group paid attention to Disaster Management services and activities. In an effort to identify which user behaviours and interactions were required to validate each of SOCIETIES' third party services, and to consider in detail the context in which the users might naturally experience each service, a high level Cognitive Walkthrough of the services was conceived and performed, by SOCIETIES partners, as a day in the life of an imagined user or persona, when the SOCIETIES Pervasive Communities platform and services were operational. The storyboards from D8.1 were used as aide memoires, as they embodied some of the user requirements, context and activites, in personas and scenarios. Demo prototypes of a few services were available to support this activity, but not all. Access to demos of the proposed services whether on phones, tablets or other server networks was userful as stories, about imagined usages and expected usages, facilitated shared understanding amongst people working on different areas and packages in the SOCIETIES project.

While the timelimits of the workshop did not allow for a detailed breakdown and cross comparison of how all of the users goals, tasks and activities and all of the proposed services, might be integrated into the targeted domain and used by students, enterprise people, and disaster management people; this co-design visioning activity helped the project researchers and developers share understandings, uncover some knowledge gaps and pose questions related to the user of the services in context, integrated together and with the SOCIETIES platform. Most importantly it was an effort to bring the people who will be participating in the trials, and for whom the services are intended, into the fore of consciousness of the project participants.

This activity was captured in feeding back the vision of a day in the life of a student, enterprise executive or disaster management worker to the entire group.

3.8 A novel variant of the Cultural Probes approach

A novel variant of traditional material Cultural Probes in combination with one of the variants Digital Cultural Probes, or Mobile Probes is proposed as a more adaptable approach which SOCIETIES will use in the case where the duration of the trials is longitudinal, and we do not want to overburden users or place them in a sense of obligation. As SOCIETIES Pervasive Communities system requires both mobile devices and humans to be part of the system, and as the organisational cultural and social norms can be expected to continue to guide user behaviour during the trials, as users will be participating within these norms present in the context of their work and study, Cultural Probes offers the potential to open up a space for alternative perspectives. As the user's input is optional, this is a risky approach, yet it is also ripe with potential.

The digital devices deployed by users in the trial will double up as tools for probing, as inspired by the example of Digital Cultural Probes, where a digital forum or Social Networking Space (if feasible using the platform, but if not using either Elgg or a more popular style of forum such as a Tumbler multi-user blog) would be set up to act as a shared public sphere for the trial participants for the duration of each trial. This will be combined with physical probes situated in the environment of the Pervasive Communities, with which participants, or their non-user colleagues, can interact.

SOCIETIES probes will be designed to evoke user responses loosely around the project's value proposition: of Discover, Connect and Organise; and the project innovations: Personalisation; Learning; User Intent; Community Preferences; Context; Community Context; Location; Relevance; Community Orchestration; Safeguarding; Trust; and Privacy. The kits will be generated from a variety of digital and material sources. Images from wider cultural sources, such as novels, films and art, in unusual settings, could be used to provoke responses. Open questions could be written on prepaid postcards and hidden around the campus, with locations provided on digital maps. Students could be invited to add colours coded feelings to digital, physical and social maps of the University. Existing social networks, such as Facebook, FourSquare, Elgg, Twitter and domain specific student computer networks, could also serve as a medium for some probes. Elements of the SOCIETIS platform could also be utilised. The exact content of the probes packets is yet to be finalised, as it is envisioned that the potential for this approach will be further informed by the pilot trials.

Including non digital probes will enable both participants and other people in the environment to respond to the concepts of Pervasive Communities, in a safe anonymous manner, beyond any suspicion of digital *surveillance*, thus extending beyond initiations of social norms existent in more formal usability tests, which might bias user responses away from expressing concerns or fears they may have about any negative aspects of the first prototypes presented.

As the natural behaviour of users using the pervasive system in the wild is of interest to researchers, introduction of the probes will have to be carefully managed and consciously controlled, so as not to disturb or bias the ecological validity of the behaviour monitoring aspect of the trial.

3.9 Measures of Usability

While we acknowledge the work done in the models for ubiquitous computing outlined in the SOTA, for clarity we revert to the ISO 9241-11 standard, which 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", although we also may reference other models in our evaluation reports. The three essential components of Usability are Effectiveness, Efficiency (or Productivity), and Satisfaction, in relation to specific user goals and specific contexts of use. Effectiveness is commonly measurable by task completion, while Efficiency is calculated by how long it took a user to complete the task. Number of errors and error recovery are also significant metrics. All these usability measurements will be possible to glean from the usage logs. Satisfaction ratings will be calculated from the SUS. The secondary community specific research questionnaires related to SOCIETIES challenges will guide more particular project decisions. More nuanced and potentially ambiguous evaluations may be obtained from Cultural Probes.

3.9.1 Devices

All aspects of devices utilized to present the SOCIETIES first prototype for trials will be documented according to the following factors:

- Device appearance (size, shape, colour, weight);
- Positioning of device (visibility, distance from observed). [69]
- User control:
- Device usage (removal, deactivation);
- Device prevention (blockage, avoidance).

4 Disaster Management First Prototype Usability Evaluation

4.1 Trial Description

The following is a brief overview of the trials in relation to the facilitation of the usability evaluation activities. The document D8.2v2 provides an in-depth description of the trials, and will be the living document, which is updated as details become cemented closer to the trials.

4.1.1 Challenges in recreating a DM scenario

As mentioned previously (see Section [69]), we face the inherent challenge in terms of recreating the full set of environmental and human factor experienced in a DM scenario. While DM experts have learnt to train and practise their skill sets, tools and procedures under realistic conditions, we do not expect the same to necessarily apply to the volunteer group. A particular case in point is the off-site volunteers and the degree to which we can motivate them to participate and the measure the impact to their emotional states. In a true DM scenario we might expect a much higher degree of motivation and also a much richer experience in terms of emotional rewards by participating. It is to be expected that this will impact the perceived and measured satisfaction and usability.

4.1.2 Location

The DM trial will have two phases, each with a different location. The first phase takes place in cyberspace. The second is staged across four locations simultaneously:

1. Neuhausen, Germany "Bundesschule Technisches Hilfswerk": This is the general school of the German disaster relief agency and the place where the off-site DM experts will reside.
2. DLR Oberpfaffenhofen or Neuhausen (on-site / outdoors): This is where the "disaster" will be enacted with a bridge or building as the main disaster area to be assessed.
3. DLR Oberpfaffenhofen or Neuhausen respectively: A selection of volunteers (on-site as well as off-site|).
4. "Cyberspace" / Virtual: A selection of off-site volunteers (i.e. people at home or at their work location).

4.1.3 Devices utilised, other webs of technologies present

Users will be provided with the following devices, corresponding to the trial location as described above:

- Standard (office) PC: Off-site volunteers at DLR Oberpfaffenhofen.
- Disaster Management Services installed on two to three notebooks: On-site assessment experts. In addition, there will be a fleet of quadrators (autonomous aerial vehicles) to assess the building or bridge, operated remotely by the volunteers. In addition, one or two augmented reality (AR) goggles will be used for overlaying critical information onto the real world view.
- SmartPhones: On-site volunteers - they will use the image geo-tagging service.
- Standard (home/office) PC: Off-site volunteers off site in "Cyberspace", (i.e. at home or in their office).
- Disaster Management Service installed on one to two notebooks: Off-site experts at Neuhausen, Germany.

4.1.4 Services

The specific third party services created for SOCIETIES Disaster Management group in response to the disaster management professionals' input in the concept development and scenario envisioning stages, as documented in D2.1 and D8.1 will be offered to trial participants for use and evaluation. Usage of each service will be logged, providing detailed information from which the effectiveness and efficiency can be deduced. A SUS survey will be used to gauge the Disaster Management users' satisfaction level with the overall SOCIETIES system experience.

We recommend that the following table be completed before the trial begins. A full description of all the services will be included in updates to D8.2v2, which is a living document and which will be updated continuously in the lead up to each of the three trials.

Disaster Management Third Party Services	**To be used by Trial Participants in 1st Prototype User Trials: Yes *or* No**	**Passed Functional and Integration Tests in WP7 by which date.**	**Comments or issues**
iDisaster Service			
IWantToHelp			
YouRNotAlone			
Disaster Data Collector			
AnalyzeThis			

Figure [69] - Disaster Management Third Party Services

4.1.5 Users, context-of-use data, and user controls

The trial will address two groups of users, at both ends of the value chain: the remote volunteers on the one hand, and the end-user professionals at the scene on the other hand.

Outline different types of users involved and describe tasks or contexts of use relevant to platform, and trial environment.

We plan the following numbers of trial participants (i.e. people playing end-user roles):

- Off-site volunteers: 10-20, possibly more
- On-site volunteers: 2-4.
- Off-site DM experts: 2-7
- On-site DM experts: 2-3

In addition, we plan that 2-3 SOCIETIES staff members be present at each of the three physical sites (i.e. on-site at Oberpfaffenhofen, off-site at Oberpfaffenhofen with the off-site volunteers, and off-site at Neuhausen) during the whole trial. This will allow some monitoring of users on a one-on-one basis.
We suggest that user preparation be 1 hour at each site.

4.1.5.1 Permissions, privacy and ethics management.

All trial participants' permissions will be sought for all data monitoring, storage and management. Privacy issues and risks will be explained to users to inform their usage of the system. User ethics will be given priority in consideration of data usage by research partners, and follow the SOCIETIES projects *ethical rules and guidelines, as defined by the consortium (WP1). [Ref]"Research_Ethics.doc"*. The ethical policies and guidelines of the host and participant organisations, namely the German Aerospace Center (DLR*)* and the German Federal Agency for Technical Relief (THW) will be followed in all of the projects interactions with trial participants.

4.1.6 Usability Methods employed.

Disaster Management trial participants will be able to provide feedback informally during the trial by communicating directly with the SOCIETIES support staff. Formally, they will be interviewed after the trial. The interview will address the same line of questions as the paper trials. The participants will be invited to fill out the short standardized SUS survey (c.f. Section 2.2.1.1 System Usability Scale (SUS)). The secondary

questionnaire related to specific issues for SOCIETIES Disaster Management applications and the Disaster Management domain, will also be given to trial participants at this time.

4.1.7 Data Collection & Analysis

The raw data of the SUS responses will be collected via paper surveys. It will be analysed according to the instructions given by Brooke and outlined in the description in Section 2.2.1.1 System Usability Scale (SUS). The SUS tests will return a result between 1 and 100. This is not a percentile score but may be converted for ease of communication. [43][41].

In addition, data related to efficacy and efficiency, related to a user's ability to complete tasks, and the time taken to do so, will be gleaned from the usage logs, the details of which will be provided closer to the trial dates.

A standard qualitative content analysis approach should be used to analyse interview data.
The usability data will be presented standard usability reports and tables similar to this one shown below will be included for brief references.

Disaster Management Third Party Services	**Usage Logs**		**SUS score**	**Comments**
	Effectiveness: Task completion	**Productivity:** Time Taken (sec)	**Satisfaction** Users subjective self-report.	
iDisaster Service			There will be just one generic SUS score for all SOCIETIES' Disaster Management Services.	
AnalyzeThis				
IWantToHelp				
YouRNotAlone				
Disaster Data Collector				

Figure 1 - Disaster Management Third Party Services Usability Table

Results from the secondary qualitative survey will be graphed where possible, and additional open answers will be analysed by a standard qualitative content analyses method, such as grounded theory. Interviews, or group discussions with participants will be recorded and treated similarly.

5 Enterprise First Prototype Usability Evaluation

5.1 Trial Description

The following is a brief overview of the trials in relation to the facilitation of the usability evaluation activities. The document D8.2 provides an in-depth description of the trials, and will be the living document, which is updated as details become cemented closer to the trials.

5.1.1 Location

The Enterprise trial will take place at a technical workshop or conference to be held in Ireland during February 2013.

5.1.2 Users

The users will be selected from delegates attending the conference. The minimum number of delegates supporting the trial will be 15, but the intention will be to engage as many delegates at the conference in the trial as we can.

5.1.3 Devices

The technical implementation of the Enterprise trial prototypes will support full deployment on Android smart phones (SOCIETIES Light Client) or any device upon which you can run the standard Java runtime (SOCIETIES Rich Client). We will also be supporting a server side deployment (SOCIETIES Cloud node) where the services will be accessible via a standard web browser. So the pervasive experience may vary depending on the user's device but the accessibility to the system will be extensive thus it will not exclude many potential users from experiencing our innovations and providing feedback.

5.1.4 Services

The specific third party services created for SOCIETIES Enterprise group in response the enterprise group input in the concept development and scenario envisioning stages, as documented in D2.1 and D8.1 will be offered to conference attendees for use and evaluation. Usage of each service will be logged, providing detailed information from which the effectiveness and efficiency can be deduced. A SUS survey will be used to gauge the Enterprise users' satisfaction level with the overall SOCIETIES system experience.

We recommend that the following table be completed before the trial begins. A full description of all the services will be included in updates to D8.2.

Enterprise Third Party Services	**To be used by Trial Participants in 1st Prototype User Trials: Yes *or* No**	**Passed Functional and Integration Tests in WP7 by which date.**	**Comments or issues**
AdHocMeetings/CollaborationTools			
NetworkingZones			
SharedCalendar / PersonalizedAgenda			
Context Aware Wall			
Conference Registration			

Figure 2 - Enterprise Third Party Services

5.1.4.1 Permissions, privacy and ethics management.

All trial participants' permissions will be sought for all data monitoring, storage and management. Privacy issues and risks will be explained to users to inform their usage of the system. User ethics will be given priority in consideration of data usage by research partners, and follow the SOCIETIES projects *ethical rules and guidelines, as defined by the consortium (WP1). [Ref]"Research_Ethics.doc"*. The ethical policies and guidelines of the host company Intel will be followed in all of the projects interactions with trial participants.

5.1.4.2 User evaluation objectives for our Enterprise Community

- *E-UO1: To investigate what levels of usability and what user interaction metaphors should be supported.*
- *E-UO2: To evaluate an easy-to-use privacy layering interface.*

5.1.5 Usability Methods employed.

Enterprise users will be able to provide feedback informally during the trial by communicating directly with the SOCIETIES support staff. Formally, they will be interviewed after the trial. The interview will address users experience of project values and innovations. The participants will be invited to fill out the short standardized SUS survey (c.f. Section 2.2.1.1 System Usability Scale (SUS)). The secondary questionnaire related to specific issues for SOCIETIES Enterprise applications and the conference domain, will also be given to trial participants at this time.

5.1.6 Data Collection & Analysis

The raw data of the SUS responses will be collected either online or via paper surveys according to which format is most suited to the Enterprise group. It will be analysed according to the instructions given by Brooke and outlined in the description in 2.2.1.1 System Usability Scale (SUS). The SUS tests will return a result between 1 and 100. This is not a percentile score but may be converted for ease of communication. [67]

In addition, data related to efficacy and efficiency, related to a user's ability to complete tasks, and the time taken to do so, will be gleaned from the usage logs.

A standard qualitative content analysis approach is recommended, to analyse interview data.

The usability data will be presented standard usability reports and tables similar to this one shown below will be included for brief references.

	Usage Logs		**SUS score**	**Comments**
	Effectiveness: Task completion	**Productivity:** Time Taken (sec)	**Satisfaction** Users subjective self-report.	
Enterprise Third Party Services				
AdHocMeetings/CollaborationTools			There will be just one generic SUS score for all SOCIETIES' Enterprise Services.	
NetworkingZones				
SharedCalendar / PersonalizedAgenda				
Context Aware Wall				
Conference Registration				

Figure 3 - Enterprise Third Party Services Usability Table

Results from the secondary qualitative survey will be graphed where possible, and additional open answers will be analysed by a standard qualitative content analyses method, such as grounded theory. Interviews, or group discussions with participants will be recorded and treated similarly.

6 Student First Prototype Usability Evaluation

6.1 Trial Description

The following is a brief overview of the trials in relation to the facilitation of the usability evaluation activities. The document D8.2 provides an in-depth description of the trials, and will be the living document, which is updated as details become cemented closer to the trials.

6.1.1 Location

The student user group trials will take place at the Heriot-Watt campus in Edinburgh, Scotland. The full description of the environment, labs and campus is available in D8.2 section 3.5 Physical Location Aspects.

6.1.2 Users

The trial will involve various groups of students from the Computer Science and Information Systems courses at Heriot- Watt. One confirmed user group consists of 30-50 Computer Science and Information Systems students who are in their 3rd year come October 2012. Their trial is expected to last for at least three months from January to March, and possibly continuing until late in the academic year, which could be as far as June 2013.

Also, more student groups may be targeted in the following academic year from October 2013. This has yet to be considered in detail, however, if it does happen, it is likely that the confirmed user group of third year students from 2012-2013 will be used. They would then be in their fourth year, where they would be undertaking dissertation projects, for which they could be encouraged to use their CSSs.

6.1.3 Devices and equipment

For the student trials, every student in a given group will receive a CSS device (e.g., a smart phone), which provides all the individual user functionality of a SOCIETIES CSS. Each group member will have the opportunity to communicate with the CSSs of fellow group members, as well as create, join, and send/receive invites for CISs.

6.1.4 Services

The specific third party services created for SOCIETIES student group in response the students input in the concept development and scenario envisioning stages, as documented in D2.1 and D8.1 will be offered to students for use and evaluation. Usage of each service will be logged, providing detailed information from which the effectiveness and efficiency can be deduced. A SUS survey will be used to gauge the users satisfaction level with the overall SOCIETIES system experience.

We recommend that the following table be completed before the trial begins. A full description of all the services will be included in updates to D8.2.

Student Third Party Services	**To be used by Trial Participants in 1st Prototype User Trials: Yes *or* No**	**Passed Functional and Integration Tests in WP7 by which date.**	**Comments or issues**
CoBrowse			
EventHerald			
CrowdTasking			
NearMe			
SocialLearningGame			
MyTV / Pesonalised Display			

Figure [67] - Student Third Party Services

6.1.5 Students as co-designers and co-developers

Tutorials and API documentation will be available for students that wish to exploit the SOCIETIES platform to create their own third party services, or 'apps' for their CSSs. Any such service will be available for sharing in a peer to peer fashion between CSSs or in a community fashion inside a CISs where multiple CSSs will be able to exploit a shared service. As the date of the trial has been moved towards the second academic term, the expectation that the 3rd year students will be able to devote time to develop SOCIETIES third party services is severely lowered due to the fact that the second academic term is a lot more demanding than the first one where the trial was originally expected to take place. However, there is a plan to include a group of 5th year Software Engineering students to design and develop third party services for the SOCIETIES platform. Capturing additional students feedback could be done though a diary method such as Experience Sampling Method (ESM) or Day Reconstruction Method (DRM), outlined in (Internal REF), but as both methods are intrusive and demanding of the students time and energy, their use could only be negotiated at the point when the final group of students is selected.

6.1.6 Cultural Probes

In addition to the SUS method, we will use a form of Cultural Probes to engage this group in further design inquiry. As the design of the SOCIETIES prototype involves complexity on so many levels, there are many new research questions and areas to investigate, but the limited duration of trials and trial participants attention to performing tests during the trials, the open ended approach of the probes will provide a counterpoint to the more empirical scientific approach, opening up the design space to accommodate multiple stories.

Inclusion of the probes could combine a mixture of mobile tasks and probing situations set up in the digital and physical environment of the trials, such as a graffiti wall, a pin and post it campus map, images of or links to cultural references of different visions of futures in film, books, art positioned in the physical or digital environment. Probes might also be a request to take part in simple community exercise or mission, which would demonstrate some system features through the participation. The mobile probes could be facilitated through Tumbler or Elgg accounts or aspects of the SOCIETIES platform, and the physical probes could be photographed at regular intervals by participants and uploaded there. The online blog could then develop into shared spaces for meta discussions around the issues raised by the new technologies, enabling and encouraging the students to express their questions and explorations around the Pervasive Communities system, in metaphor, image or text in an undirected and unrestrictive way. When invitations to share responses to initial prototypes for D8.1 were issued, few users chose to comment online. This may indicate an unwillingness to use Elgg or simply that the people who participated in the earlier trials did not have ongoing issues related to the technologies to discuss. Whilst we may expect that participants in a live trial will have much to share, we should also facilitate multiple ways for users to post responses to digital probes and make participation as light and easy as possible.

"Again, it's important to balance the ecological validity of gathering user data in situ against the interruption of everyday activity flow that recording personal observations causes."

We will endeavour to introduce the probes in such a way so as to preserve the ecological validity of the behavioural monitoring aspect of the trial.

6.1.7 Objectives for Usability Testing for First Student Prototype:

A broad range of objectives for the user trials for SOCIETIES first prototype is listed in D8.2. A subset of those goals specific to usability is listed here.

3.2.2 Specific User evaluation objectives for our Student Community

- *S-UO1: To investigate what levels of usability and what user interaction metaphors should be supported.*
- *S-UO2: To evaluate an easy-to-use privacy-layering interface.*
- *S-UO3: To evaluate the end-user experience and user-acceptance of the presented system.*

6.1.7.1 Permissions, privacy and ethics, and data management.

Students' permissions will be sought for all data monitoring, storage and management. Privacy issues and risks will be explained to users to inform their usage of the system. User ethics will be given priority in

consideration of data usage by research partners, and follow the SOCOETIES projects *ethical rules and guidelines, as defined by the consortium (WP1). [Ref]"Research_Ethics.doc"*. The ethical policies and guidelines of the University will be followed in all of the projects interactions with students.

6.1.8 Methods to be employed.

The System Usability Scale (SUS) standardised user evaluation survey (c.f. Section 2.2.1.1 System Usability Scale (SUS)) will be issued to all participants, and they will be asked to complete it. A secondary questionnaire related to specific issues for SOCIETIES Student applications and the university domain, drafted from the research questions identified in D8.2, and reproduced in Annex A, Section 1 Qualitative User Feedback from Students will also be given to trial participants at this time. The ESM or DRM method may be employed by a small number of users, if some students are sufficiently motivated to participate in a diary study. Cultural Probes will also be utilised as design inquiry method, to enhance understanding, reflection, and communication between researchers and student trial participants.

6.1.9 Data Collection, Analysis and Interpretation

The raw data of the SUS responses will be collected either online or via paper surveys according to which format is most suited to the user groups.

It will be analysed according to the instructions given by Brooke and outlined in the description in section 2.2.1.1 System Usability Scale (SUS). The SUS tests will return a result between 1 and 100. This is not a percentile score but may be converted for ease of communication. [33]

In addition the data related to efficacy and efficiency, related to a user's ability to complete tasks, and the time taken to do so, will be gleaned from the usage logs.

The usability data will be presented standard usability reports and tables similar to this one shown below will be included for brief references.

	Usage Logs		**SUS score**	**Comments**
Student Third Party Services	**Effectiveness:** Task completion	**Productivity:** Time Taken (sec)	**Satisfaction** Users subjective self-report.	
CoBrowse			There will be just one generic SUS score for all SOCIETIES' Student Services.	
EventHerald				
CrowdTasking				
NearMe				
SocialLearningGame				
MyTV / Pesonalised Display				

Figure 4 - Student Third Party Services Usability Table

Results from the secondary qualitative survey will be graphed where possible, and additional open answers will be analysed by a standard qualitative content analyses method, such as grounded theory. Interviews, or group discussions with participants will be recorded and treated similarly.

Where Cultural Probes are utilised, the treatment and use of the material generated does not have to be formally analysed, as it is not an epistemological method. The results of Cultural Probes are valued not for their ability to close down the design field, but are valuable for provoking reflective process, inspiring further design enhancement, and allowing for multiplicity of interpretations to be valid. A description of the probes and the activities they inspired accompanied by a set of photographs or outtakes from the material generated could be presented to document the process. In addition the material will provide the basis for participatory workshops at the end of the trials.

7 Method of feedback of usability results to the project

The System Usability Scale results will be calculated to a single score for each of SOCIETIES user groups. This will provide a benchmark against which future prototypes might be assessed.

The on-device logging activities will provide a wealth of data from which we will glean information in relation to the efficiencies and effectiveness of applications for each domain.

The charts included in this deliverable will be completed and results will be circulated to all people working on the SOCIETIES project.

The secondary questionnaires for each user group generated from specific research questions identified in D8.2 will provide focused and useful guidance to aspects of users' expectations, and requirements of interest to SOCIETIS. Results will be graphed where possible, and additional open answers will be analysed by a standard qualitative content analyses method, such as grounded theory.

Interviews conducted on site with participants in the case of Disaster Management and Enterprise will be parsed and analysed according to standard content analyses and grounded theory methods, and findings, which emerge, will be recorded in descriptive reports.

Any additional material generated by further qualitative research, such as Cultural Probes, which cannot be included in an end of trial debriefing workshop, will be kept for future project workshops, to which if possible the students, enterprise workers, and disaster management workers, would be invited to participate. Activities and processes related to Cultural Probes will be documented live in a shared Tumbler or Elgg account and interesting aspects of this content, and related discussions will be presented in a descriptive report, including photographs, text and images.

8 Conclusions

This document D8.4 presents the Specification of usability testing for first prototype, which is an element or subset of exercises and texts to be organised concurrently with the full set of user trials described in D8.2.

In this document we have presented a State of the Art on usability evaluation for pervasive and ubiquitous systems, beginning with a look back at how usability evaluation approaches have evolved in UCD design and development projects in general, before going on to consider some of the issues particularly pertinent to evaluation of ubiquitous computing systems. We also refer to the literature to investigate four qualitative approaches with very different flavours: Experimental Sampling Method (ESM), Day Reconstruction Method (DRM), System Usability Scale standardised survey, and Cultural Probes. The first three are self-reporting methods, whereas the Cultural Probes are a design enquiry approach.

The second section of this document framed the Specification of Usability Evaluation for SOCIETIES First Prototype Trial with reference to several pertinent issues particular to the project. We described the project workshop in Edinburgh where significant decisions, such as which usability survey to use, and shared realisations occurred; such as where some third party application design decisions inadequately match some daily rituals of potential users.

The methodology chosen for user evaluation for the trials is to use the System Usability Scale (SUS) survey for each of the trials, so as to have a benchmark usability score, both for longitudinal trials within each group, and as a set of results comparable amongst Students, Enterprise executives, and Disaster Management experts. Acknowledging the limitations of what the SUS can measure, a secondary qualitative survey is also to be derived based on the specific questions generated for each community in D8.2 and reproduced in the Annex A of this document. Further additional qualitative approaches might also be included where feasible. However, which approaches might be best suited to each trial will have to be negotiated with the trial managers and participants in each particular case.

We are sensitive to the balance required between allowing users time to engage with the services, participate in the other tests planned for the project, (as outlined in D8.2), and putting a burden, social obligation on users to respond, or any extra workload on project partners in collection or analysis tasks. In fact, it is expected that the user trials will provide the project with an abundance of data about user behaviour, from on device logging, which can also be gleaned for usability indicators of efficiency and efficacy. In light of these factors, we suggest that a variant of Cultural Probes, including digital and non-digital probes, could be an interesting secondary approach for SOCIETIES user trials. While not an evaluation method, with results and analysis that fit into scientific method, it is a hermeneutic approach that values ambiguity, and offers an opportunity to strengthen relationships, and inspire valuable insights for both developers and users. Most importantly the Cultural Probes could create a safe space for multiple interpretations and expressions about the project value propositions and innovations, offering a chance for alternative more expansive SOCIETIES visions to exist alongside those primary project Prototypes.

Details regarding the trial planning, participants, location, devices, data capture and analyses are presented in the previous sections: Section 6 the Student group, Section 5 for Enterprise group, and section 4 for the Disaster Management group.

While there are no definitive usability methods, which can allow for all the variables and causal relationships to be discovered, in relation to how people will interact with and use the project's first prototype from which the usability can be clearly established, use of a standardised usability test will allow for some generalisation of results. Secondary questionnaire will be needed to capture user input as required for more specific formative purposes, for the SOCIETIES project. On the other hand, tacit information could be gained via participatory approaches, such as Cultural Probes distributed in the digital and physical environments, in which the trials will take place, could facilitate engaging users with some of the more abstract and ambiguous aspects of the proposed technologies, potentially leading to a better understanding between trial participants and project researchers, and supporting more active user engagement with the evaluation activities planned for SOCIETIES final prototype, to be outlined in D8.6 Specification of usability testing for final prototype. We have barely begun to discuss the further layer of complexity that evaluations for social factors, imply when considering the socio technical model and the popularity of mobile social computing in conjunction with pervasive systems. Yet it is clear that that the network effect and community considerations

will guide the next research directions for usability of SOCIETIES Pervasive Communities applications in future user evaluations for the final prototype D8.6 Specification of usability testing for final prototype.

In conclusion, usability as a field is challenged to keep up with developments in computing, and there are no clear standards or guidelines, widely adopted, to advise on the particular and difficult problems of how to evaluate pervasive computing applications. Usability evaluation for complex systems is not clearly defined, and shares ground with user experience, and user research, as researchers apply mixed approaches in efforts to engage and include potential end users. Pervasive and social systems are incomplete until the users arrive.

Field trials are becoming *de rigour* for mobile and pervasive applications, as complex environmental, social, mobile, and context-informed interactions are required to facilitate human behaviours and generate the data required to fuel and demonstrate system functions. Yet they are limited when it comes to uncovering causal relationships, and novel approaches to quasi experiments may be required. The qualities of reliability and validity in usability evaluation assume that it is possible to repeat tests, to get similar results, and generalise results for a broader group of similar users. However, how will this be resolved with the highly personalized and context based responsive and unpredictable experiences that people have in the social and pervasive environments of the SOCIETIES system is not evident; where there is no specific fixed context of use, where the users' preferences and attention fluctuate according to the context of use, and users goals' are complex shifting patterns of intentions, social negotiations and desires.

9 References

[1] J. Dumas, "The great leap forward: The birth of the usability profession (1988-1993)," *Journal of Usability Studies*, vol. 2, no. 2, pp. 54–60, 2007.

[2] G. Cockton, "Usability Evaluation." Aarhus, Denmark: The Interaction Design Foundation., 2012.

[3] A. Dillon, "Beyond usability: process, outcome and affect in human-computer interactions," 2002.

[4] H. Petrie and N. Bevan, "The evaluation of accessibility, usability and user experience," *The Universal Access Handbook*, pp. 20–1, 2009.

[5] N. Bevan, "International standards for usability should be more widely used," *Journal of Usability studies*, vol. 4, no. 3, pp. 106–113, 2009.

[6] S. Bødker, "When second wave HCI meets third wave challenges," in *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles*, 2006, pp. 1–8.

[7] M. J. Muller, "Participatory design: the third space in HCI," *Human-Computer Interaction: Development Process*, p. 165, 2009.

[8] M. Hassenzahl, M. Soegaard, and R. F. Dam, "User Experience and Experience Design.," in *Encyclopedia of Human-Computer Interaction*, 2011.

[9] M. Hassenzahl and N. Tractinsky, "User experience - a research agenda," *Behaviour & Information Technology*, vol. 25, no. 2, pp. 91–97, Mar. 2006.

[10] J. Nielsen, "10 Heuristics for User Interface Design." [Online]. Available: http://www.useit.com/papers/heuristic/heuristic_list.html. [Accessed: 15-Aug-2012].

[11] W. D. Gray and M. C. Salzman, "Damaged merchandise? A review of experiments that compare usability evaluation methods," *Human-computer interaction*, vol. 13, no. 3, pp. 203–261, 1998.

[12] C. Wharton, J. Rieman, C. Lewis, and P. Polson, "The cognitive walkthrough method: A practitioner's guide," in *Usability inspection methods*, J. Nielsen and R. L. Mack, Eds. New York: Wiley, 1994.

[13] S. Krug, *Don't make me think! : a common sense approach to Web usability*. Indianapolis, Ind.: Que, 2000.

[14] S. Viller and I. Sommerville, "Social analysis in the requirements engineering process: from ethnography to method," in *Requirements Engineering, 1999. Proceedings. IEEE International Symposium on*, 1999, pp. 6–13.

[15] K. H. Madsen and P. H. Aiken, "Experiences using cooperative interactive storyboard prototyping," *Communications of the ACM*, vol. 36, no. 6, pp. 57–64, 1993.

[16] A. Agarwal and A. Meyer, "Beyond usability," 2009, p. 2919.

[17] P. Wright and J. McCarthy, "Empathy and experience in HCI," in *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, 2008, pp. 637–646.

[18] E. B. N. Sanders, "From User-Centered to Participatory Design Approaches." .

[19] J. McCarthy and P. Wright, "Technology as experience," *interactions*, vol. 11, no. 5, pp. 42–43, 2004.

[20] Brian Whitworth and Rob Friedman, "Reinventing academic publishing online. Part II: A socio-technical vision," *First Monday; Volume 14, Number 9 - 7 September 2009*, 2009.

[21] E. B. N. Sanders and P. J. Stappers, "Taylor & Francis Online :: Co-creation and the new landscapes of design - CoDesign - Volume 4, Issue 1." [Online]. Available: http://www.tandfonline.com/doi/full/10.1080/15710880701875068#h12. [Accessed: 21-Aug-2012].

[22] S. Greenberg and B. Buxton, "Usability evaluation considered harmful (some of the time)," 2008, p. 111.

[23] I. Mulder and P. J. Stappers, "Co-creating in Practice: Results and Challenges," in *15th International Conference on Concurrent Engineering (ICE 2009)*, 2009, pp. 22–24.

[24] H. J. Kim, J. K. Choi, and Y. Ji, "Usability Evaluation Framework for Ubiquitous Computing Device," in *2008 Third International Conference on Convergence and Hybrid Information Technology*, Busan, Korea, 2008, pp. 164–170.

[25] S. Consolvo and M. Walker, "Using the experience sampling method to evaluate ubicomp applications," *IEEE Pervasive Computing*, vol. 2, no. 2, pp. 24–31, Apr. 2003.

[26] L. Downey, "Group Usability Testing: Evolution in Usability Techniques," 2007, vol. Vol. 2, Issue 3, May 2007,, pp. pp. 133–144.

[27] K. M. Scott, "FEATUREIs usability obsolete?," *interactions*, vol. 16, no. 3, p. 6, May 2009.

[28] K. Connelly, "On developing a technology acceptance model for pervasive computing," in *9th International Conference on Ubiquitous Computing (UBICOMP)-Workshop of Ubiquitous System Evaluation (USE), Springer, Innsbruck, Austria*, 2007, p. 520.

[29]"Kim et al. - 2008 - Usability Evaluation Framework for Ubiquitous Comp.pdf." .
[30] B. Poppinga, M. Pielot, N. Henze, and S. Boll, "Unsupervised User Observation in the App Store: Experiences with the Sensor-based Evaluation of a Mobile Pedestrian Navigation Application," *observing the mobile user experience*, p. 41, 2010.
[31] S. Carter, J. Mankoff, S. Klemmer, and T. Matthews, "Exiting the Cleanroom: On Ecological Validity and Ubiquitous Computing," *Human-Comp. Interaction*, vol. 23, no. 1, pp. 47–99, Jan. 2008.
[32] A. Oulasvirta, "Fielding Usability Evaluation," *IEEE Pervasive Computing*, 2011.
[33] T. R. Hansen, J. E. Bardram, and M. Soegaard, "Moving Out of the Lab: Deploying Pervasive Technologies in a Hospital," *IEEE Pervasive Computing*, vol. 5, no. 3, pp. 24–31, Jul. 2006.
[34] J. Redish, "Expanding usability testing to evaluate complex systems," *Journal of Usability Studies*, vol. 2, no. 3, pp. 102–111, 2007.
[35] M. Langheinrich, "Privacy in ubiquitous computing," *Ubiqutious Computing Fundamentals*, pp. 96–156, 2009.
[36] F. Zhu, S. Carpenter, and A. Kulkarni, "Understanding identity exposure in pervasive computing environments," *Pervasive and Mobile Computing*, Jul. 2011.
[37] S. Moran, "User perceptions of system attributes in ubiquitous monitoring : toward a model of behavioural intention," The University of Reading, 2011.
[38] M. Dodge and R. Kitchin, "'Outlines of a world coming into existence': pervasive computing and the ethics of forgetting," *Environment and Planning B: Planning and Design*, vol. 34, no. 3, pp. 431–445, 2007.
[39] M. Weiser, "The computer for the 21st century," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 3, no. 3, pp. 3–11, 1999.
[40]"Mark Weiser - Wikipedia, the free encyclopedia." [Online]. Available: http://en.wikipedia.org/wiki/Mark_Weiser. [Accessed: 17-Aug-2012].
[41] J. Scholtz and S. Consolvo, "Toward a framework for evaluating ubiquitous computing applications," *Pervasive Computing, IEEE*, vol. 3, no. 2, pp. 82–88, 2004.
[42] S. S. Intille, J. Rondoni, C. Kukla, I. Ancona, and L. Bao, "A context-aware experience sampling tool," in *CHI'03 extended abstracts on Human factors in computing systems*, 2003, pp. 972–973.
[43] T. RiisgaardHansen, J. E. Bardram, and M. Soegaard, "Interactive hospital project."
[44] S. Carter, J. Mankoff, and J. Heer, "Momento: support for situated ubicomp experimentation," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2007, pp. 125–134.
[45] J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay, "MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones," in *Proceedings of the 5th international conference on Mobile systems, applications and services*, 2007, pp. 57–70.
[46] D. A. Norman, *Living with complexity*. Cambridge, Mass.: MIT Press, 2011.
[47] M. Scheermesser, H. Kosow, A. Rashid, and C. Holtmann, "User acceptance of pervasive computing in healthcare: Main findings of two case studies," 2008, pp. 205–213.
[48] S. Moran and K. Nakata, "Taylor & Francis Online :: Analysing the factors affecting users in intelligent pervasive spaces - Intelligent Buildings International - Volume 2, Issue 1." [Online]. Available: http://www.tandfonline.com/doi/abs/10.3763/inbi.2009.0040. [Accessed: 14-Aug-2012].
[49] J. Kirakowski, "SUMI Background Reading." [Online]. Available: http://sumi.ucc.ie/sumipapp.html. [Accessed: 14-Aug-2012].
[50] J. Sauro, "Measuring Usability: Quantitative Usability, Statistics and Six-Sigma." [Online]. Available: http://www.measuringusability.com/. [Accessed: 19-Aug-2012].
[51] J. Brooke, "SUS-A quick and dirty usability scale," in *Usability evaluation in industry*, P. Jordan, B. Thomas, B. Weerdmeester, and I. McClelland, Eds. Taylor and Francis.
[52] Y. S. Ryu, "Mobile Phone Usability Questionnaire (MPUQ) and Automated Usability Evaluation," in *Human-Computer Interaction. New Trends*, vol. 5610, J. A. Jacko, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 349–351.
[53] J. Sauro, "Measuring Usability with the System Usability Scale (SUS): Measuring Usability." [Online]. Available: http://www.measuringusability.com/sus.php. [Accessed: 09-Aug-2012].
[54] T. S. Tullis and J. N. Stetson, "A Comparison of Questionnaires for Assessing Website Usability." [Online]. Available: http://www.usabilityprofessionals.org/usability_resources/conference/2004/UPA-2004-TullisStetson.pdf. [Accessed: 14-Aug-2012].
[55]"A New Formula for Quantitative UX Decision Making | UX Magazine." [Online]. Available: http://uxmag.com/articles/a-new-formula-for-quantitative-ux-decision-making. [Accessed: 17-Aug-2012].

[56] M. Csikszentmihalyi and R. Larson, "The Experience Sampling Method.," *New Directions for Methodology of Social & Behavioral Science*, vol. No 15., no. Mar 1983, pp. 41–56.
[57] D. Kahneman, "A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method," *Science*, vol. 306, no. 5702, pp. 1776–1780, Dec. 2004.
[58] E. Karapanos, J. Zimmerman, J. Forlizzi, and J. B. Martens, "User experience over time: an initial framework," in *Proceedings of the 27th international conference on Human factors in computing systems*, 2009, pp. 729–738.
[59] K. Boehner, J. Vertesi, P. Sengers, and P. Dourish, "How HCI interprets the probes," 2007, p. 1077.
[60] B. Gaver, T. Dunne, and E. Pacenti, "Design: Cultural probes," *interactions*, vol. 6, no. 1, pp. 21–29, Jan. 1999.
[61] T. Mattelmäki, *Design Probes*, vol. 0. University of Art and Design Helsinki, 2006.
[62] W. W. Gaver, A. Boucher, S. Pennington, and B. Walker, "Cultural probes and the value of uncertainty," *interactions*, vol. 11, no. 5, p. 53, Sep. 2004.
[63]"Interview with Bill Gaver | Design Research Network." [Online]. Available: https://www.designresearchnetwork.org/drn/content/interview-bill-gaver. [Accessed: 14-Aug-2012].
[64] O. S. Iversen and C. Nielsen, "Using digital cultural probes in design with children," in *Proceeding of the 2003 conference on Interaction design and children*, 2003.
[65] S. Hulkko, T. Mattelmäki, K. Virtanen, and T. Keinonen, "Mobile probes," in *Proceedings of the third Nordic conference on Human-computer interaction*, 2004, pp. 43–51.
[66] J. Bardram and A. Friday, "Ubiquitous computing systems," *Ubiquitous Computing Fundamentals*, pp. 37–94, 2010.
[67]"WAMMI - Questionnaire." [Online]. Available: http://www.wammi.com/questionnaire.html. [Accessed: 16-Aug-2012].
[68] D. M. Russell, "The Streitz perspective: computation is ubiquitous, yet must be designed for human use (a festschrift for Norbert Streitz)," *Universal Access in the Information Society*, vol. 11, no. 3, pp. 233–238, Sep. 2011.
[69] G. D. Abowd, G. R. Hayes, G. Iachello, J. A. Kientz, S. N. Patel, M. M. Stevens, and K. N. Truong, "Prototypes and Paratypes: Designing Mobile and Ubiquitous Computing Applications," *IEEE Pervasive Computing*, vol. 4, no. 4, pp. 67–73, Oct. 2005.
[70]"SUMI Questionnaire Homepage." [Online]. Available: http://sumi.ucc.ie/index.html. [Accessed: 31-Aug-2012].

Annex A

A.1 Formal Performance Metrics and Feedback for WP4, 5 & 6

The following test tables, are taken from D8.2, where we have taken the objectives described in Section 4.2 and the analysis of the evaluation points set by WP4 and WP5 (see Appendix A.2). The tests are listed in order of priority. The first column gives the unique test identifier. The second column gives the test prioritisation. The next column gives a summary of the test description. The last four columns attempt to capture more detailed requirements for each of these tests, as they relate particularly to WP4-6.

Test ID	Priority	Test description	Input	WP4 Relation	WP5 Relation	WP6 Relation
D1	H	This test should be conducted at the beginning of the trial. We plan to obtain the volunteer and expert members' opinion on community services, especially automatic community join and community driven assistance for DM.	Question-naire.			
D2	H	This test aims to get users' opinion on the perceived benefit/risks of automatic management of communities of rescuers based on expertise / availability / workload / previous collaborations / human-social preferences (additionally location / task / institution, country / language / sex / etc.). Compared to manual management or semi-manual.	Observation, question-naire.		Proactivity, Context Inference	User interface relating proactivity / autom. Mgt. to the service itself
D3	H	Test whether DM volunteer users would like services to start automatically on their behalf, such as automatic recommendation that their expertise would now be of direct value.	Observation, question-naire.		Proactivity, Context Inference	User interface relating proactivity to the service itself
D4	H	Test whether volunteers and experts would like to join or even be automatically added into sub-communities based on shared criteria.	Observation, question-naire.			

Test ID	Priority	Test description	Input	WP4 Relation	WP5 Relation	WP6 Relation
D5	H	Test how the users trust the decision making mechanism provided by SOCIETIES, i.e. should the system ask users for confirm first before any action?	Observation, question-naire.		Proactivity, Context Inference	User interface relating proactivity to the service itself
D6	H	Test whether users trust the privacy control provided by SOCIETIES	Question-naire.		Privacy Mgt.	
D7	H	Test how useful it is to share access to physical resources (e.g. sensor controls, cameras, robotic arms, UAV controls, etc.) in online communities	TBD, logging; Observation, question-naire.		Sensor Mgt. Context Inference	Service efficiency, HMI
D8	H	Test how useful it is to share information collected from physical sensors (typically those that collect user related information such as location, activity) in online communities	TBD, logging; Observation, question-naire.		Sensor Mgt. Context Inference	Service may range from fully automated DM expert allocation, to just providing support for manual allocation
D8	H	Test if it is useful to provide access to external services (e.g. translation services) in online communities	Comparison with ground truth information (e.g. translation accuracy, speed)			Service efficiency, ease of use
D9	M	Test which are the best ways of setting up a pervasive communication infrastructure for sharing physical resources.	TBD	Network resources		
D10	M	Test whether value is provided by advanced context inference and personalisation.	TBD		Context Inference and Personal-isation	
D11	M	Test whether transition between connected and disconnected operations worked smoothly from a user perspective	TBD, Observation, Logging	Networking		
D12	L	Test whether users would like to be aware of actions or contexts of other community members, and whether such information would affect their own decisions.	Observation, question-naire.			

Test ID	Priority	Test description	Input	WP4 Relation	WP5 Relation	WP6 Relation
D13	L	Test whether the volunteers would like to join a random volunteer community without specifying any particular interest.	Observation, question-naire ; logging			How do services that are related to a community (e.g. translation) present themselves to the user?
D14	L	Test whether knowing about other community members would affect volunteers' willingness to join the community.	Observation, question-naire.			

A.2 Qualitative User Feedback from Students

This table is taken from D8.2 and includes the qualitative tests that will be carried out on the student group with more qualitative techniques such as questionnaires and focus groups.

Test ID	Evaluation Question	Output	Methodology plus interference, controls issues if relevant	WP4 / WP5 Relation	Objective Reference
S77	Do users feel shared services are useful?	transcript	Questionnaire / Group discussion		S-RO5 S-RO7
S78	Do users feel that they shared more personal information than they first expected they would?	transcript	Questionnaire / Group discussion	• Privacy & Trust	
S79	Did the users get annoyed with the volume of pop-ups and notifications during the trial?	transcript	Questionnaire / Group discussion	• User Agent	S-UO3
S80	Did the users feel they trusted the system to make automatic decisions/behaviours more as the trial progressed?	transcript	Questionnaire / Group discussion	• User Agent	S-TO2
S81	Do the users feel that most recommended communities were relevant to them?	transcript	Questionnaire / Group discussion	• Intelligent Community Orchestration	S-RO5 S-RO7
S82	Do the users feel that most suggested automatic behaviours were appropriate?	transcript	Questionnaire / Group discussion	• User Agent	
S83	What criteria do users feel was most influential in driving them to join or reject a recommended community?	transcript	Questionnaire / Group discussion	• Intelligent Community Orchestration	
S84	How user friendly did the users find the SOCIETIES configuration screens?	transcript	Questionnaire / Group discussion		S-UO1 S-UO3
S85	Were there any configuration functions that the users felt could be added/improved?	transcript	Questionnaire / Group discussion		S-UO1 S-UO3

S86	Did the users feel that the smartphone was an appropriate CSS device?	transcript	Questionnaire / Group discussion		S-UO1 S-UO3 S-TO1 S-TO2 S-TO3
S87	What limitations did the users find with using a smartphone as a CSS devices?	transcript	Questionnaire / Group discussion		S-UO1 S-UO3 S-TO1 S-TO2 S-TO3
S88	Did the users feel that the SOCIETIES system was useful to them as students?	transcript	Questionnaire / Group discussion		S-RO10 S-UO3 S-UO1
S89	How did 3p service developers find the development and deployment process?	transcript	Questionnaire / Group discussion		S-UO1 S-UO3
S90	Why did users create CISs?	transcript	Questionnaire / Group discussion	• CIS Mgmt	S-RO3 S-RO10 S-UO1 S-UO3
S91	How do users feel about wearing sensors?	transcript	Questionnaire / Group discussion		
S92	How do users feel about being monitored?	transcript	Questionnaire / Group discussion	• User Agent	S-RO7
S93	Do users trust that the system will protect their privacy? If yes, why?	transcript	Questionnaire / Group discussion	• Privacy & Trust	S-RO7
S94	Do users understand the concept of privacy preferences?	transcript	Questionnaire / Group discussion	• Privacy & Trust	S-RO7
S95	Do users understand the concept of data obfuscation?	transcript	Questionnaire / Group discussion	• Privacy & Trust	S-RO7
S96	Would users like a tool that would warn them that they are about to share media containing sensitive information to inappropriate users or user groups?	transcript	Questionnaire / Group discussion	• Privacy & Trust	S-RO7
S97	Would it be helpful for end users to have services fitted with a clear visual estimation of trust?	transcript	Questionnaire / Group discussion	• Privacy & Trust	S-RO7

S98	Would it be helpful for end users to have other users fitted with a clear visual estimation of trust?	transcript	Questionnaire / Group discussion	• Privacy & Trust	S-RO7
S99	Are there any types of automatic behaviour that users favoured?	transcript	Questionnaire / Group discussion	• User Agent	S-RO4 S-RO5 S-UO1 S-UO3
S100	Are there any types of automatic behaviour that users did not like?	transcript	Questionnaire / Group discussion	• User Agent	S-RO4 S-RO5 S-UO1 S-UO3
S101	Did the users ever go along with an automatic behaviour suggestion, even though they didn't really want it?	transcript	Questionnaire / Group discussion	• User Agent	S-RO4 S-RO5 S-UO1 S-UO3
S102	Were users happy with the pop-up mechanisms? How could they be improved?	transcript	Questionnaire / Group discussion	• User Agent	S-UO3

A.3 Qualitative User Feedback from Enterprise Community

This table, which has been copied from D8.2, includes the qualitative tests that will be carried out on the enterprise community with more qualitative measuring tools.

Test ID	Evaluation Question	Output	Methodology plus interference, controls issues if relevant	WP4 / WP5 Relation	Objective Reference
E75	Do users feel shared services are useful?	transcript	Questionnaire		E-RO5 E-RO7
E76	Do users feel that they shared more personal information than they first expected they would?	transcript	Questionnaire	• Privacy & Trust	
E77	Did the users get annoyed with the volume of pop-ups and notifications during the trial?	transcript	Questionnaire	• User Agent	E-UO3
E78	Did the users feel they trusted the system to make automatic decisions/behaviours more as the trial progressed?	transcript	Questionnaire	• User Agent	E-TO2
E79	Do the users feel that most recommended communities were relevant to them?	transcript	Questionnaire	• Intelligent Community Orchestration	E-RO5 E-RO7
E80	Do the users feel that most suggested automatic behaviours were appropriate?	transcript	Questionnaire	• User Agent	
E81	What criteria do users feel was most influential in driving them to join or reject a recommended community?	transcript	Questionnaire	• Intelligent Community Orchestration	
E82	How user friendly did the users find the SOCIETIES configuration screens?	transcript	Questionnaire		E-UO1 E-UO3
E83	Were there any configuration functions that the users felt could be added/improved?	transcript	Questionnaire		E-UO1 E-UO3
E84	Did the users feel that the smartphone was an appropriate CSS device?	transcript	Questionnaire		E-UO1 E-UO3 E-TO1 E-TO2 E-TO3
E85	What limitations did the users find with using a smartphone as a	transcript	Questionnaire		E-UO1

	CSS devices?				E-UO3 E-TO1 E-TO2 E-TO3
E86	Did the users feel that the SOCIETIES system was useful to them?	transcript	Questionnaire		E-RO10 E-UO3 E-UO1
E87	Why did users create CISs?	transcript	Questionnaire	• CIS Mgmt	E-RO3 E-RO10 E-UO1 E-UO3
E88	How do users feel about wearing sensors?	transcript	Questionnaire		
E89	How do users feel about being monitored?	transcript	Questionnaire	• User Agent	E-RO7
E90	Do users trust that the system will protect their privacy? If yes, why?	transcript	Questionnaire	• Privacy & Trust	E-RO7
E91	Do users understand the concept of privacy preferences?	transcript	Questionnaire	• Privacy & Trust	E-RO7
E92	Do users understand the concept of data obfuscation?	transcript	Questionnaire	• Privacy & Trust	E-RO7
E93	Would users like a tool that would warn them that they are about to share media containing sensitive information to inappropriate users or user groups?	transcript	Questionnaire	• Privacy & Trust	E-RO7
E94	Would it be helpful for end users to have services fitted with a clear visual estimation of trust?	transcript	Questionnaire	• Privacy & Trust	E-RO7
E95	Would it be helpful for end users to have other users fitted with a clear visual estimation of trust?	transcript	Questionnaire	• Privacy & Trust	E-RO7
E96	Are there any types of automatic behaviour that users favoured?	transcript	Questionnaire	• User Agent	E-RO4 E-RO5 E-UO1 E-UO3

E97	Are there any types of automatic behaviour that users did not like?	transcript	Questionnaire	• User Agent	E-RO4 E-RO5 E-UO1 E-UO3
E98	Did the users ever go along with an automatic behaviour suggestion, even though they didn't really want it?	transcript	Questionnaire	• User Agent	E-RO4 E-RO5 E-UO1 E-UO3
E99	Were users happy with the pop-up mechanisms? How could they be improved?	transcript	Questionnaire	• User Agent	E-UO3