WeKnowIt

Emerging, Collective Intelligence for Personal, Organisational and Social Use

FP7-215453

D7.6.2

Final consumer and emergency response use case evaluation protocols

Dissemination level	Public
Contractual date of delivery	Month 33, 31-12-2010
Actual date of delivery	25-02-2011
Workpackage	WP7, Case studies
Task	T7.2.3, User trial and Evaluation
Type	Report
Approval Status	Approved
Version	1.13
Number of pages	41
Filename	D7.6.2_2011-02-25_v1.13 (TID)

Abstract:

This deliverable describes both consumer social group and emergency response case evaluation protocols that will be used for the evaluation of the first WKI prototype.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

co-funded by the European Union

History

Version	Date	Reason	Responsible
0.1	24/11/2010	TOC	TID
1.1	17/12/2010	CSG Mobile guidance protocol description	TID
1.2	20/12/2010	ER Evaluation methodology	USFD
1.3	10/01/2011	Contribution to CSG Mobile and Athens trials' description	VOD
1.4	13/01/2011	New structure for Athens and Krakow trials	TID
1.4b	13/01/2011	Athens trials description	VOD
1.5	13/01/2011	Crakow trials description	SMIND
1.6	24/01/2011	CSG Pre-travel and post-travel protocol description	YAHOO!
1.8	03/02/2011	Integration	TID
1.9	07/02/2011	Review	EM-KA
1.10	08/02/2011	Corrections	YAHOO!
1.11	23/02/2011	Corrections	USFD
1.12	25/02/2011	Integration	TID

Author list

Organization	Name	Contact Information	Reason
TID	Manuel Escriche	mev@tid.es	Author
USFD	Vitaveska Lanfranchi	v.lanfranchi@dcs.shef.ac.uk	Author
VOD	Costis Kontopoulos	costis.kontopoulos@vodafone.com	Author
SMIND	Tomek Kaczanowski	tomasz.kaczanowski@softwaremind.pl	Author
YAHOO!	Borkur Sigurbjornsson	borkur@yahoo-inc.com	Author
EM-KA	Andreas Sonnenbichler	andreas.sonnenbichler@kit.edu	Reviewer

Executive Summary

WeKnowIt is exploring the concept of Collective Intelligence – a form of intelligence emergent from the cooperation of multiple layers of intelligence using a shared knowledge base. These Intelligent layers provide a set of services used by a set of tools developed for the two project scenarios: Emergency Response and Consumer Social Group.

This deliverable describes the evaluation protocols to be used during the final evaluation trials in the two scenarios ER and CSG of the WeKnowIt Project.

The protocol used in the Emergency Response scenario uses a user-centric approach, where users are involved during the development phase and a final evaluation with a research approach including a longitudinal study and a standard usability evaluation.

The protocol used in the Consumer Social Group scenario is focused on knowing the end user satisfaction by getting feedback on usage dimensions. The pre-travel and post-travel tools use standard usability questionnaires and the mobile guidance application ad-hoc questionnaires.

The evaluation trials are run in different settings by different project partners, i.e. ER trials will run in Sheffield by USFD and SCC and in Krakow by SMIND, while CSG trials will run in Madrid by TID, in Athens by VOD and CERTH-NTUA, and in Krakow by SMIND.

Abbreviations and Acronyms

API	Application Programming Interface
CSG	Consumer Social Group
E-WKI	The WeKnowIt System as seen by the users
ER	Emergency Response
FLO	Forward Liaison Officer
GPS	Global Positioning System
PC	Personal Computer
POI	Point of Interest
REST	Representational state transfer
SCC	Sheffield City Council
SUS	Usability and Satisfaction Questionnaires
UK	United Kingdom
UI	User Interface
USFD	The University of Sheffield
TID	Telefónica Investigación y Desarrollo
WKI	WeKnowIt
WP	Work Package
XML	Extensible Markup Language

Table of Contents

List of Figures

1. Introduction

This deliverable describes the evaluation protocols for both case studies: Emergency Response and Consumer Social Group. This deliverable collects preparatory information to execute tasks T7.2.3 Evaluation and user trial, as framed in WP7, and according to section B.1.3.6 Work Package description of WeKnowIt Annex I – "Description of Work."

This deliverable is an improved release over deliverable D7.6.1 "Initial Consumer and Emergency Response Use Case Evaluation Protocols".

The main objective of WeKnowIt project is to develop novel techniques for exploiting multiple layers of intelligence from user-contributed content, which together constitute Collective Intelligence, a form of intelligence that emerges from the collaboration and competition among many individuals, and that seemingly has a mind of its own.

To this end, WeKnowIt project has chosen two different but complementary case studies to demonstrate the wide applicability of its technologies and research activities: Emergency Response and Consumer Social Group.

This deliverable's objective is to describe the evaluation protocol, i.e. the actions or steps taken to perform the end users evaluation. This protocol excludes actions taken to realize the technical validation, which is considered satisfactory.

Structure of the Document

The deliverable has been organized in the following way: after presenting the overall approach in this **first section**, the **second section** presents different approaches to evaluating interactive systems.

In the **third and fourth sections**, we present the protocols for the evaluation of both studied cases (ER and CSG). While the Emergency Response protocol is presented in section 3, section 4 is dedicated to the definition of the Consumer Social Group protocol.

Finally, **section five** summarizes the deliverable content.

2. Evaluation Methodology – State of the Art

This section gives an outlook by describing different approaches to evaluating interactive systems [2] and continues by describing the options for participatory evaluation. Non-participatory evaluation techniques are briefly described followed by user-centric evaluations and the various measurement techniques which can be employed.

2.1. Non-Participatory Evaluations

Several formal evaluations can occur during the development lifecycle and need not necessarily to involve potential users of the system. Evaluations of this form can focus purely on the system itself or they can use domain experts to identify usability problems with any stage of the interface design.

2.1.1. Functional Evaluation

Functional evaluations are undertaken without regard for the system interfaces - the purpose of the functional evaluation is to determine if the system components are meeting the requirements laid out in the specification. This process can occur at different levels of granularity. At *high levels* of granularity the evaluation could ask abstract questions about the system as a whole. For example, a functional evaluation can be used to determine if the system performs the required computations when a certain part of the interface is accessed. At *lower levels* of granularity unit testing can be employed to evaluate the system at a component or function level. *Unit testing code* is typically written in parallel to production code and takes the form of automated tests of individual functions with each module. Thus the tests ensure that each part of each module conforms to the specification.

2.1.2. Cognitive Walkthrough

A *cognitive walkthrough* is carried out by an expert with reference to a particular task or tasks that the system should carry out. Typically this involves an expert running through the same procedure that a user may follow when using the system and considering if the interface is *clear and consistent at each step*. Whilst this does not account for the variability found when testing with the target user base the process can identify UI problems early on in development. The walkthrough can act as a pre-test, ensuring that the system reacts as one would expect given the system specification.

This technique can also be performed as participatory evaluation by involving users that will run through the system assessing the functionalities, highlighting issues and providing comments.

This technique has been heavily adopted for the Emergency Response case study: cognitive walkthrough of the interface were performed at regular intervals both by expert developers and ER Sheffield Team members, to make sure the interface was respecting the user needs, that it was executing the task in an easy and efficient way and that the overall user experience was satisfying.

During cognitive walkthrough sessions the previously identified requirements were revised and new requirements were added, that were later validated during interviews and focus groups.

2.1.3. Heuristic Evaluation

A *heuristic evaluation* utilises a set of rules or 'heuristics' and attempts to apply these rules to the interface. Typically such rules address generic core user interface issues. Such heuristics can, like cognitive walkthrough, take place at any stage during the interface design and can be carried out by a single evaluator. For example, Nielsen's heuristics [3] is a list of 10 guidelines a system should follow including having clear functionality and not relying on the user to remember how to carry out tasks. These heuristics act as high level guidelines for what users typically expect from complex systems and how those systems should react and present information.

2.2. Participatory Evaluation

Participatory evaluation is a method of evaluation in which external parties are used to judge the effectiveness of the system. As with the non-participatory evaluations this need not only occur when there is a fully working system and can take place at any stage of the development lifecycle. For example, in order to assess the effectiveness of the interface design, paper mock-ups of the interface can be shown to the user and the user can 'interact' with the system by simply stating what actions they would perform with the paper interface. The evaluator then acts as the system and presents the correct paper interface for each corresponding action. Furthermore parts of the system can be replicated by having external bodies fulfil the functionality as in a Wizard-of-Oz style evaluation [8]. For example, an unfinished speech recognition module could be replaced by manual transcription for the purposes of evaluation.

Generally, participatory evaluations can be split into whether they involve the system being evaluated in the environment in which it will be used (*field trials*) or if the system is evaluated in a controlled environment, designed to mimic the real environment as much as possible (*lab studies*).

2.2.1. Field Trials

In *field trials* the system is evaluated as if it were in use and largely free of experimenter control. In WKI, for example, this would involve providing a set of users with suitable equipment to access the WKI system and

these users would then either be free to use the system as they see fit or a testing scenario could be developed in which the users 'act' out an example scenario which makes use of the WKI system. Field trials are designed to assess the system in the ways it would be used in reality and as such it is harder to control for variance in the way that the system is used. Because of this, the evaluation has more relevance since the experimenters are able to see exactly how the system reacts in the real world. However, because the system is being used freely and outside of the experimenter's domain it is difficult to ask specific questions about how the interface and system is working.

For example the WeKnowIt project recently participated in a simulation of an emergency exercise, in cooperation with the Sheffield City Council and Sheffield Emergency Services. The simulation was of a chemical attack on a major building in the city centre. The emergency services were testing how effectively they could set up decontamination tents and process as many members of the public and sufficiently decontaminate them.

WeKnowIt participated in two capacities. First of all, a member of the team observed the entire incident from the beginning to dealing with decontaminated people in the Sheffield United football ground. The purpose of this observation was to collect a large set of geo-located images for use in future experiments and to test how well the WeKnowIt ER demonstrator could display and process.

In addition to this, five citizen participants were asked to arrive at the incident at a random time and to capture some images using a stripped down version of the WeKnowIt interface to examine how a typical user annotates emergency information in the field.

2.2.2. A longitudinal studies

Long-term investigations have been adopted in many disciplines, but only recently they started being applied to usability [14], introducing the concept of longitudinal usability evaluation.

Longitudinal usability studies are emerging as a new methodology to evaluate complex applications over time, as they allow focusing on longer-term usability issues that may affect user performance and satisfaction for a given task.

Standard usability evaluation methodologies tend to focus on short term first time evaluations, biasing the "results more towards "discoverability" or "learnability" problems [10]. Whilst these approaches could be sufficient in simple domains, where tasks are straightforward, complex domains and applications require a longer term perspective, to make sure a new application is really helping the user and improving the way their tasks are executed. Moreover when dealing with Intelligent Applications, it becomes of fundamental importance to study the long-term effects of an applications, to understand if intelligent, adaptive technologies have a

different degree of perceived usability over time. This could be due to the fact that some complex features require time to be learnt and time to become fully functional so to make a difference in the user experience.

User experience has been defined in ISO FDIS 9241-210 as:

"A person's perceptions and responses that result from the use and/or anticipated use of a product, system or service", while usability has been defined in ISO FDIS 9241-210 "Extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use." As pointed out by Bevan [2] both these definitions lack a time dimension. The definition of usability is lacking in considering how changes happen over time, while the definition of user experience does not take into account how the user experience is affected by the actual interaction with a product/system over time [13].

For example an adaptive interface based on automatic machine-learning needs time to develop models and features that can then be exploited in the user interaction. A usability test of such an application can be successful if the application is previously trained [5][12][7] but does not represent an effective study of how the users use and benefit from the technology, as the application would be tailored beforehand by expert users.

What happens instead when a complex application is given to users in a complex domain? How long before the intelligent features become apparent and help the users instead of hindering or causing frustration? Will the application be perceived more useful after 3 months of usage that after 1 hour? What are the added benefits? What are the longer-term usability issues that should be fixed? And how does the use of the application change the task the users are performing?

Longitudinal studies often adopt a methodology was based on triangulation [3][4][14], using multiple methods to collect data to better cover the multiple aspects of a complex domain. This is to make sure that as many aspects of the domain as possible are covered, even if each method will give just a partial insight on the domain.

2.2.3. Lab Studies

In *lab studies* the system is evaluated in a controlled environment in a regimented fashion. The advantage of lab studies is that because the experimenter has more control over the environment, unwanted sources of variance can be removed or dampened - so any results generated from the evaluation are more relevant, albeit at the cost of being less realistic. A further advantage of lab studies over field trials is that the system need not exist as a functional whole in order for the evaluation to occur - different modules of the system can be assessed independently of each other.

Furthermore, since there is less of a requirement for the system to be fully interactive, lab studies can also evaluate paper prototypes and make use of Wizard-of-Oz procedures [8] to simulate non-functional parts of the system. In this way the evaluator is able to select the parts of the system or interface that they wish to focus on and simulate the non-functional or non-existent parts of the system.

2.2.4. Usability Evaluation

Usability evaluations aim to verify the usability of the system with real users in terms of

- Effectiveness
- Efficiency
- User satisfaction

ISO 9241-11 suggests that measures of usability should cover

- effectiveness (the ability of users to complete tasks using the system, and the quality of the output of those tasks),
- efficiency (the level of resource consumed in performing tasks)
- user satisfaction (users' subjective reactions to using the system)

These measures can be calculated using different approaches: for example measures of effectiveness can be the types of task that are carried out and how effectively they are reached. This could be measured by system logs, time taken for each task but also by asking specific questions in a questionnaire or in a user interview. More details on measurement techniques are provided in Section 2.3.

2.3. Measurement Techniques

Within both field trials and lab studies there are a number of methods of measuring and recording the interaction between the participant and the system. The principal methods are listed below, although it is often the case that these approaches are combined in the resulting evaluation in order to fully capture and evaluate the interaction.

2.3.1. Think Aloud

Think aloud encourages the user to both actively report and reflect on the activity they are currently involved in. This means that the evaluator is able to assess not only what the participant is doing, but also why they are doing it, and in some cases why they are not doing certain actions. A drawback to this approach is that the user is drawn out of reacting naturally to the interface which may mean that their responses are not entirely naturalistic. In addition, it is an unnatural process that may require some prompting from the evaluator and this may also affect the users reaction to the system and any process measurements that are made during the experiment.

2.3.2. Video Recording

An alternative option is to video record the user interaction with the system. Video recording can be superior to screen recording (see below) as it allows the experimenter to see both the interface and users response to the interface. Video recording can suffer as the evaluator needs to ensure that the viewing angle is correct and that the relevant information is being captured as part of the recording process. There is also a lengthy post-hoc analysis process that the evaluator must employ in order to extract relevant information from the video recording.

2.3.3. Logging

In addition to screen recording, logging can also be used to measure the interaction between the user and the system. Logging allows the evaluator to generate reasonably precise measures of how the user interacts with the system and the time between discrete actions. As with screen recording logging can only evaluate what happens internally in the system. Logging has the advantage that the extraction process is often quicker as there is no coding required by the evaluator. A drawback of logging is that it is often challenging to strike a balance between gathering enough information to make the log worthwhile but not capturing so much information that it is difficult to extract useful information.

2.3.4. Eye Tracking

With eye tracking, the subject sits in front of a screen which has been augmented with an eye tracking device which is able to determine with some degree of accuracy where on the screen the subject is currently looking. This data can be analysed separately to determine which parts of the interface the user focuses on when interacting with the system. This approach has the advantage of being able to determine which parts of the interface are most salient to the user and whether key parts of the interface are being overlooked.

2.3.5. Screen Recording

Here, software is installed onto the target machine which allows the evaluator to make a full-screen recording whilst the evaluation takes place. This can then be analysed and coded once the evaluation is complete. A drawback to this approach is that only the interaction is recorded and so it can be difficult to locate when the user has trouble understanding some part of the interface or the underlying reason for this confusion. The recording can, however, be supplemented with an audio recording of the participant performing think aloud in order to address this problem.

2.3.6. Physiological Recording

It is also possible to record the physiological reactions of the user to the interface using lightweight sensors. This allows the evaluator to assess the physical response that the user has to the interface and, in turn, identify interfaces or points within the interaction that increases the level of stress that the user has or confuses them. However, it can be difficult to accurately interpret the meaning of physiological recordings and is not suitable for all types of interface.

2.4. Post Hoc Measurement Techniques

In addition to measuring the interaction between the user and system, it is also of importance to question the user about the system after the interaction is complete. This allows the user to interact with the system without any distractions whilst the evaluator takes measurements regarding the interactions. The evaluator can then tune the post-hoc measurements towards the user interaction. For example, the session may reveal that some part of the interface was troublesome for the user and the post-hoc measurements can address this specific part of the interface.

2.4.1. Questionnaires

Questionnaires can consist of Likert [9] questions (a statement which the user should express to what extent they agree or disagree with) and questions which require longer answers. Questionnaires should be relatively short in order to maximise the legitimacy of the responses and to make the best use of the time in the evaluation session.

2.4.2. Interviews

Interviews that take place after the session can also be used to directly ask the user questions about the interface. Typically, a small number of questions are prepared in advance which then act as a means of exploring the interface with the assistance of the user. Interviews can be used to probe particular actions that occurred in the evaluation session or to examine the questionnaire responses in more detail. Interviews can also be a useful method of forcing the user to expand on their thoughts about the interface.

2.4.3. Focus Groups

Focus groups are participative evaluation methods focus to elicit qualitative opinion, perceptions and attitudes from users towards an interface or a system. Typically the group setting is interactive, with some basic questions asked by the experimenter and free discussion between the groups' members. For example regular focus groups were held for the Emergency Response case study with members of the ER Teams in

Sheffield and Doncaster, to assess the new version of the prototype, identify further needs and requirements.

3. Emergency Response Case Study

The Emergency Response Case Study will be evaluated to understand how the final prototype would help users in dealing with a scenario like the one presented in D7.1. The scenario was set in Sheffield on the 25th of June 2007. Some Sheffield citizens are at their workplace in the Sheffield City Council when the water starts to rise in the River Dom and one of them is caught up in the floods while going back home from his workplace. A group of Emergency Responders based in the Sheffield City Council and Forward Liason Officers (FLO) on the emergency scene send and receive information about the emergency and use it to support the decision making process. The scenario also follows the citizens whilst trying to gather information about the emergency.

The final evaluation for the ER prototype will be split in three parts to best cover all the sides of the ER Scenario and the type of users that could be involved in an emergency:

- A longitudinal study of the ER applications for organisational intelligence, involving ER professionals to demonstrate how the emergency responders could use the application in real world emergencies and how the application would support their task.
- Evaluation of standard usability for personal intelligence, involving citizens, to demonstrate how citizens could help by uploading information and could access information about an emergency using the only system.
- A simulation of emergency, with Emergency Response and personnel participating to the exercise, to simulate how the ER application could be used in an emergency both from ER team members to upload and access information and by citizens to contribute information to the Emergency Responders.

In the following sections we will describe the three parts of the final evaluation, defining the methodology and the evaluation criteria.

3.1. Longitudinal Study for ER Professionals

The longitudinal study for the WeKnowIt project aims to understand how the user experience of an intelligent application (i.e. an application that contains several modules, running in the background, that extract information and use it to present a multimodal personalised interface over collective intelligence to the users) in a complex domain is affected by time, especially how the interaction changes when the users become expert and how this affects user effectiveness and satisfaction.

In particular some basic research questions that we aim to answer are:

- How is the user experience of intelligent applications affected by changes over time?

- How long does it take to users to accept an intelligent system and have benefits from its usage?
- Does the way users interact with an intelligent system changes over time and what are the factors that impact on it?
- Is the task affected by the use of an intelligent application? If yes, how and after how long?

3.1.1. Participants

The participants are members of Sheffield and Doncaster Emergency Response team. They will be chosen with a high degree of familiarity with the field and with new technologies. The subjects will not be compensated for the study, as they will undertake it as part of their normal duties in the emergency response team.

The participants will use the intelligent application in a complementary way to other methodologies, to make sure no data will be lost because of system malfunctions as this could not be risked in such a real-life complex domain.

The total number of participants will be split into two different categories, Forward Liason Officers (FLO, Emergency Response Personnel that is always on the go, collecting information around the city and being sent to emergency areas when needed) and office-based personnel.

3.1.2. Methods

The methods adopted are:

1. Long term pilot study
2. Field Observation
3. Focus Group Evaluation
4. Regular interviews
5. Data analysis

The methods will interact as follows: FLOs will be given a smart-phone to run the WKI application when out to gather information. Their task is to take pictures, upload and tag them, using the assisted tagging capabilities. Office-based personnel will be notified whenever new content is added to the system and will be reviewing the content at least every 2 days.

A questionnaire will be distributed regularly and repeatedly to all the participants, aiming to ascertain the effectiveness, efficacy and user satisfaction over time.

Follow-up calls will be scheduled regularly and repeatedly with each user, to collect qualitative impressions and suggestions. Field observation will also be performed, shadowing FLOs and office-based personnel to better understand how their task is influenced by the use of the application.

All the actions the users will perform on mobile phones and on the website will be logged, and analysed regularly before the field observation days and the telephone calls.

3.1.3. Measures

Usability will be evaluated using the ISO standard metrics of effectiveness, efficacy and user satisfaction.

Quantitative measures will be collected such as the time taken for executing tasks, as each action is time-stamped in the system logs.

Qualitative measures will be collected during the investigation such as observer's impressions and users comments.

3.2. Evaluation of standard usability for personal intelligence

This usability evaluations aims to understand if the ER application is suitable to support citizens in an emergency by allowing them to upload information and access information.

3.2.1. Methodology

The adopted methodology will be based on triangulation [3][4][14], using multiple methods to collect data to better cover the multiple aspects of such a complex domain.

As opposed to the evaluation for organisational intelligence, this study will be short-term as posting information about an emergency is not a normal task for a citizen so no analysis about the task and the long-term effects of the application can be done.

3.2.2. Participants

The participants that will be recruited for the evaluation will be English speakers, with a mixed degree of familiarity with new technologies: the aim is to evaluate the usability of the interface for every type of person, from the most experienced with new technologies to the less experienced, as an emergency response system should be usable by anyone in an emergency circumstance.

3.2.3. Methods

The methods adopted will be

- Task execution with observer
- Usability questionnaire
- Interviews

Before starting the evaluation a task will be chosen, representative of a real work task: the task will be used to test the main functionalities of the system and will act as guidance to the user. The task will be presented in

written form and will be executed under the observation of the experimenter that will also be available to answer any question from the participant.

In order to collect demographic data, a small demographic questionnaire will be given to the participant to fill before the session starts, including questions on gender, age, job and familiarity with new technologies.

The answers to these questions will then be used in the analysis phase to identify correlations between any of the demographic variables and the usability of the system.

The questionnaire adopted will be very simple, aiming to answer basic usability questions, such as the SUS questionnaire, a simple usability scale for a global assessment of systems.

The questionnaire will be based on a *Likert scale* [9], where the participant has to indicate the degree of agreement or disagreement with the statement on a 5 (or 7) point scale.

An example of questions that will be present in the questionnaire is:

- I thought the system was easy to use
- I think that I would need the support of a technical person to be able to use this system
- I found the various functions in this system were well integrated

The questions will be formulated with regards to specific aspects of the system for example:

1. Overall reaction to the system

1	2	3	4	5
difficult				easy

1	2	3	4	5
frustrating				satisfying

1	2	3	4	5
dull				stimulating

1.I am able to complete my work quickly using this system.

1	2	3	4	5
Strongly agree				Strongly disagree

2.I feel comfortable using this system.

1	2	3	4	5
Strongly agree				Strongly disagree

3. It is easy to find the information I need.

1	2	3	4	5
Strongly agree				Strongly disagree

4. The information provided with the system is easy to understand.

1	2	3	4	5
Strongly agree				Strongly disagree

5. The information is effective in helping me complete my work.

1	2	3	4	5
Strongly agree				Strongly disagree

6. The interface of this system is pleasant

1	2	3	4	5
Strongly agree				Strongly disagree

7. I like using the interface of this system.

1	2	3	4	5
Strongly agree				Strongly disagree

8. This system has all the functions and capabilities I expect it to have.

1	2	3	4	5
Strongly agree				Strongly disagree

2. Learning

1. Learning to operate the system

1	2	3	4	5
difficult				easy

2. Exploring new features by trial and error…… ……not applicable ☐

1	2	3	4	5
difficult				easy

3. Performing tasks is straightforward

1	2	3	4	5
never				always

3. Accessing information

1. Navigating the map interface

1	2	3	4	5
difficult				easy

1	2	3	4	5
Not useful at all				helpful

2. Browsing the photos was:

1	2	3	4	5
difficult				easy

1	2	3	4	5
Not useful at all				helpful

3. Browsing the incident forms was:

1	2	3	4	5
difficult				easy

1	2	3	4	5
Not useful at all				helpful

4. The time filtering widgets was……not applicable ☐

1	2	3	4	5
difficult				easy

1	2	3	4	5
Not useful at all				helpful

5. The tag filtering widgets was……not applicable ☐

1	2	3	4	5
difficult				easy

1	2	3	4	5
Not useful at all				useful

6. The tag analytic visualisation was……not applicable ☐

1	2	3	4	5
difficult				easy

1	2	3	4	5
Not useful at all				useful

7. The time analytic visualisation was……not applicable ☐

1	2	3	4	5
difficult				easy

1	2	3	4	5
Not useful at all				useful

4. System Capabilities

1. System speed ☐

1	2	3	4	5
low				high

During the interviews the experimenter will be able to ask questions that require a more articulate answer and get opinions and comments from the participants. These questions can be generic or targeted to evaluate a specific functionality

An example of questions that could be asked is:

1. Should you describe what ER Application allows you to do to a colleague what would you say?
2. How do you feel about the level of control you had?
3. What is your opinion on the query composition, i.e. selecting and specifying terms?

3.3. ER and Citizens Exercise

A full day workshop simulating an emergency has been organised in cooperation with Sheffield City Council to evaluate the WKI research.

During the full day there will be a simulation of an emergency and two user groups (citizens and ER personnel) will use the WKI system to capture, store and browse data. At the end of the day the ER personnel will take part to a post-incident exercise to investigate the emergency and the lessons learnt.

3.3.1. Intelligence Upload and Access

A comparative study, comparing current practices with the new system introduced by WKI, will be undertaken. In order to do so, the exercise will be split into two phases:

- Data collection and analysis using traditional practices
  - Telephone and personal messages
  - Pen and paper
- Data collection and analysis using WKI
  - Remote messaging
  - Remote content upload
  - Content Analysis

The study focuses on analysing the data produced by the 2 user groups:

- how different/similar they are
- what is the focus on each user group
- is there any trend/pattern that can be associated to the user group or to any other variable?

The two user groups will then be asked to review and comments on the data collected by the other group, the type of data, the annotations etc.

3.3.2. Post-Incident Management Exercise

After the simulation exercise the ER personnel will be involved in a post-incident management exercise, using WKI tools to review the collected data and perform incident investigation and lessons learnt analysis.

This study will allow to evaluate the WKI Post-Incident Management tool but also to understand how the ER professional use the collected data and if there is any difference in usage between the data collected by citizens and ER personnel and to what variable they are linked (different annotations, different level of granularity etc.)

4. Consumer Social Group Case

This section describes the trials protocols for the different phases of the CSG scenario: Pre-travel and Post-travel and mobile guidance; and finally the two additional evaluation trials, which will cover the complete scenario, to be run in Athens and Krakow.

4.1. Pre-travel

The WeKnowIt pre-travel prototype was evaluated in the first round of evaluation using both WeKnowIt consortium members and external users [18]. Based on the outcome of that evaluation improvements have been made to the prototype. The improvements were mainly in terms of improved usability. Furthermore we have integrated several WeKnowIt services that have become available since the evaluation of the first version of the prototype.

For the second round of evaluation for the pre-travel prototype, we will make use of the evaluation initiatives taking place in Athens and Krakow (See Section 4.4). The subjects will use the pre-travel prototype to plan a trip in the historical centers of Athens and Krakow, respectively. The usability of the prototype will be measured using standard usability questionnaire such as the SUS questionnaire. In addition we will ask questions about the usability and usefulness of specific interface components and integrated technologies. Last, we will have open questions where the participants can give feedback on further improvements to the prototype (see Section 2.4.1 on methodology).

Using the data gathered in this evaluation round we can compare the usability scores of the two rounds of prototypes and aggregated the open feedback to collect insights about further improvement options.

4.2. Post-travel

The post-travel prototype, Fannr, was developed after the first evaluation round of the WeKnowIt prototypes. The second round of evaluation for the WeKnowIt project will thus serve as a pilot evaluation of the prototype. As the prototype is applied to annotating Flickr photos the participants will need to fulfill several criteria: 1) They need to be active Flckr users; 2) They need to be used to annotate their Flickr photos with either tags or geo-tags. Due to these constraints we have chosen not to evaluate it together with the pre-travel prototype.

The evaluation will be a field trial where the subject will use the application out of control of the experimenter (see Section 2.2.1). We will assemble a group of active Flickr users that fulfill the conditions stated above, and ask them to annotate some of their own Flickr photos using the Fannr prototype. The evaluation will take place remotely, meaning that the participants will use their own computers at home or work.

After they have used the prototype for a while, we will ask the participants to fill in questionnaires about their experience using the Fannr prototype and their attitude toward photo annotations in general (see Section 2.4.1). The questionnaires will be in four parts:

1. **Fannr usability**: A questionnaire where we ask the users to describe their experience with the Fannr prototoype. We will ask questions about specific parts of the Fannr interface. E.g., Did you find the grouping of similar tags useful? The goal of this part of the survey will be to evaluate directly the implementation of the Fannr prototype.
2. **Tagging functionality**: A questionnaire where we ask the users about their attitude toward different tagging functionalities. These questions will be abstracted away from the particular implementation of tagging functionalities in Fannr. E.g., Do you find it useful to group similar tag-suggestions together? The purpose of this part of the survey will be to compare the attitude toward certain tagging functionalities and their implementation in Fannr (Part 1). I.e., if a functionality gets a low score in Part 1 but a higher one in Part 2 it indicates that the functionality may be desired but its implementation in Fannr needs improvement.
3. **Tagging behavior**: A questionnaire about how people annotate their photos. These questions will be about tagging behavior in general, such as what types of tags users find important, whether they find capitalization or phrases important, and at what granularity level they want to geo-code their photos (exact, city, or country). The goal of this part of the survey is to get a better understanding of users' needs for tagging systems. This information can be used to refine the functionality of the Fannr prototype.
4. **Flickr usage**: A questionnaire about how people use Flickr. Such as how many photos they have, how many contacts, etc. Using the results of this questionnaire we can look at the results of the previous questionnaires for certain sub-groups, such as, frequent users and occasional users.

All the questionnaires above will have questions where the participant is asked to mark their agreement with a certain statement using a 5-value Likert scale. In addition, in the Fannr usability part there will be an open question where the participants will be able to give additional feedback regarding the Fannr prototype.

In addition to the actions of the participants will be logged (see Section 2.3.3) in order to produce a summary of the performance of different WeKnowIt services integrated into the prototype.

4.3. Mobile guidance

The mobile guidance protocol to evaluate the Mobile Guidance application has been structured in three stages: pre-evaluation, evaluation and post evaluation. Each stage needs different attitude and has differentiating features.

Figure 1: CSG – Mobile - Evaluation protocol stages

The pre-evaluation stage requires thinking ahead and getting everything ready: evaluators, evaluand, scenarios and criteria. The evaluation stage requires the evaluators to experience the evaluand, which is the object of the evaluation, within a definite scenario so that value is determined and expressed in the questionnaires. Finally, the post-evaluation stage requires data analysis and drawing conclusions and recommendations.

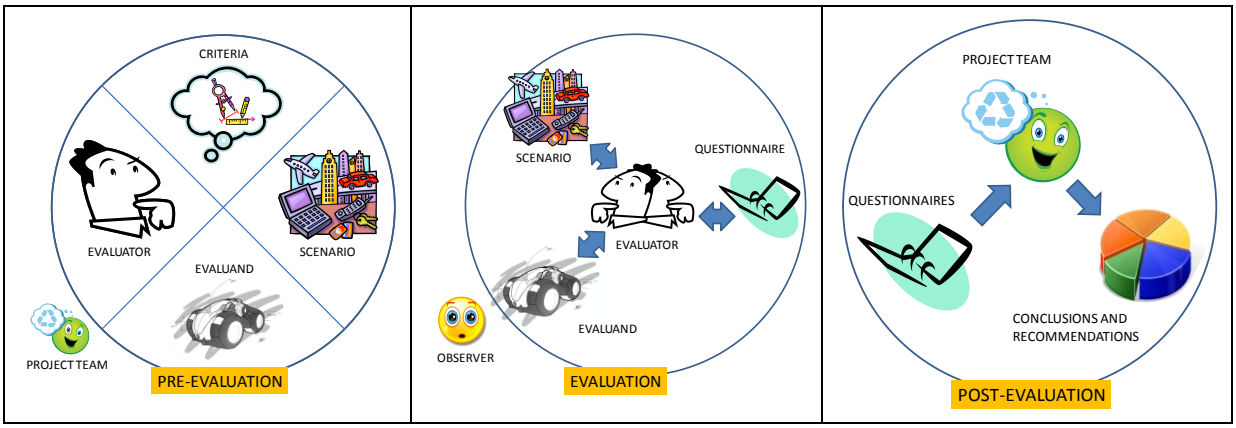

Figure 2: CSG – Mobile - View of evaluation protocol stages

Following sub-sections further explain the details involved in each stage.

4.3.1. Pre-evaluation

During this stage the project team identifies all elements needed to run the evaluation and execute all tasks to have them ready. It is a preparation activity.

There are four main elements needed: evaluators, evaluand, scenario and criteria.

Evaluators are the people who will experience the evaluand in the context provided by the scenarios and will determine the value provided by applying defined criteria.

Evaluand is the object of the evaluation. Scenario is the context where the evaluators

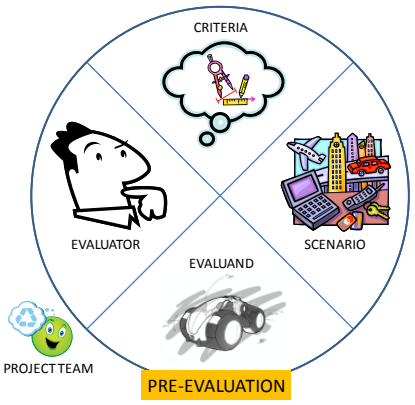

Figure 3: CSG – Mobile – Pre-evaluation

will interact and experience the evaluand. Finally, Criteria are the aspects or dimensions of interest over the evaluand's features such as usability, complexity, etc which allow evaluators apply their understanding and express value.

In the mobile guidance application, these four elements are outlined next:

Evaluators

Since WeKnowIt project is running the last iteration, Evaluators will be groups of end users with no relationship with the project team so that no conflict of interest is possible.

Evaluators will be end users accustomed to using mobile phones; it is preferable they have used their phones for more advanced functionality than just phone call.

Evaluators should enjoy the idea of exploring the touristic routes/scenarios just supported by the mobile phone.

There is no restriction to age, gender, formation or cultural background.

Evaluand

The evaluand, which is the object of the evaluation, will be the mobile guidance application running on mobile phones: iPhone and Android.

Current version sports improvements in many aspects: URL access, map control, menu, appearance space optimization and integrated services: ClustTour, Contextual and Personal POI recommendations, Group messaging.

The complete description will be provided in the corresponding deliverable D7.4.2 Consumers Social Group Case Study Implementation.

Figure 4: CSG – Mobile - Evaluand

Scenarios

TID evaluation will take place in Madrid, in the same area used for the first round, since it is a historic touristic area with many points of interest. Information has been provided in previous reports.

Additionally two project partners, VOD and SMIND, agreed at consortium level to run evaluations in other historic scenarios with many enjoyable points of interest: Athens and Krakow.

Next section is devoted to these two additional evaluations and allows these partners to provide details and particularities for their own evaluation settings.

Figure 5: CSG – Mobile - Madrid

As a result three different partners: VOD, SMIND and TID will run evaluations based on the same protocol, being each one in charge of the evaluation running in its own scenario.

Criteria

The aspects of interest driving evaluators' attention over the evaluand's features are the dimensions to appreciate and assess value.

While the user-evaluator is experiencing the evaluand in its context, which is the scenario, not every aspect or event is equally important. The evaluation criteria points out what aspects, what behaviours, what features from the evaluand deserve attention because they create value for the end user.

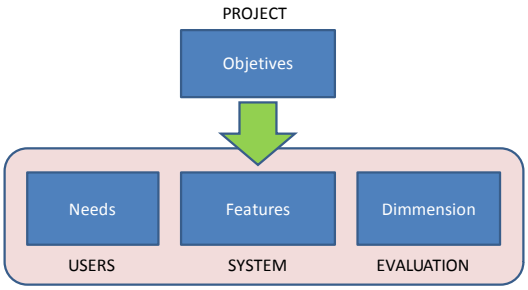

Figure 6: CSG – Mobile - Evaluation scope

The evaluation is oriented towards the end user's needs; therefore it should not be biased by the project objectives.

The different dimensions of assessment can be grouped into general levels according to their research scope. The following table gives an overview of the assessment dimensions which are selected for the second round evaluation of the CSG case study.

Level	Dimensions
Application Access	Availability Responsiveness
Application Interaction	Usability Appearance
User Needs	Usefulness Suitability Completion
User Preferences	Most interesting features/functionalities Overall usage experience

Following paragraphs elaborate on the previous terms:

Application Access Level:

This level of investigation examines the degree at which the system is operable and committable to the user requested tasks.

Owing to the wireless data connection of the application, the perceived notion of accessibility is much dependent and therefore impacted by common deficiencies innate to the nature of cellular telecommunications

(e.g. network coverage holes, handovers to low-bandwidth radio access technologies, etc.). Therefore, prior to the evaluations, it should be noticed to the participants to be aware of the network coverage while judging access level.

Measurement dimensions:

- **Availability**: by this term we determine the number of times that all functional items were indeed reachable, functioning and able to be used
- **Responsiveness**: it refers to the ability of the functional units to complete the assigned tasks within a given time. This dimension tracks the user's perception about "delays" in getting back a response in any of his/her operations.

Application Interaction Level:

This level of investigation assesses the qualitative attributes about the ease-of-use and aesthetics of the user-interface.

Measurement dimensions:

- **Usability**: Under this dimension fall the aspects of the learning curve and efficiency about the user interface. The learning curve expresses the level at which the appearance of the application self-explanatorily and intuitively guides the first-time user in performing the tasks. On the other hand, efficiency in the design of the interface greatly affects the ease and speed of performing a requested function. Usability can often be tracked via complexity, which is a reverse metric of the former.
- **Appearance**: The way the application is designed and appears in front of the user, its aesthetics can contribute to the overall usage satisfaction.

User Needs Level:

At this level we investigate the degree at which the application fulfils the users' needs or expectations, and therefore it is highly relative to the user context, circumstances and understanding. The CSG evaluation scenario allows fixing them by providing touristic scenarios with focused activities.

Measurement dimensions:

- **Usefulness**: it gives the level of satisfaction or benefit an individual gains from utilizing or consuming the provided service. This dimension can give an indication of the user's intention of use.
- **Suitability**: it assesses the degree the systems sport the right qualities for its intended purpose; i.e. according to the user's requirement and expectations in the target scenarios.
- **Completion**: it identifies the degree to which the user feels that nothing needs to be added to the system, since it has the complete set of features and functionalities expected for the intended scenario.

User Preference Level:

Finally, at this scope we try to identify the aspects and features of the application that have mostly attracted the interest and preference of the users. Furthermore, we will record how their overall experience was from using the application.

Summary

Dimensions in **the access level** explore whether the mobile guidance application is available and responsive. This level is basic; it makes next level possible.

Dimensions in **the interaction level** explore how the exchange between user-evaluator and evaluand takes place: appearance and usability. This is an intermediate level; it makes next level possible.

Dimensions in **the needs level** explore how the user's need level are satisfied in the target scenario: usefulness, suitability and completion. This is final level; if previous levels are not well-fulfilled, assessment of these dimensions is not possible.

Dimensions in **the user preference level** explore most attracting features and overall experience with the application.

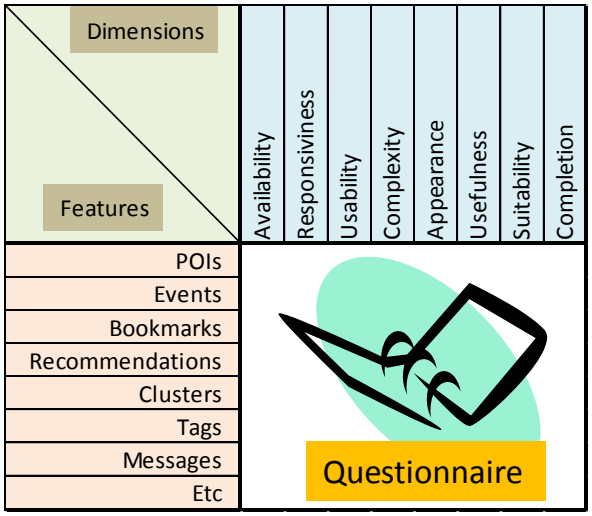

Figure 7: CSG – Mobile - Questionnaire scope

Evaluation criteria take the final form of questionnaires. By crossing features and dimensions we identify definite questions. Once this step is over, numeric scales where to express the value appreciated are included appropriately in the forms. This way the questionnaire gets ready for the evaluation.

4.3.2. Evaluation

In this stage the project team turns into an observer and the evaluators become main actors: they use the evaluand (the guidance mobile application) in its real context (scenario) with an attitude for appreciating the value contributed by the evaluand to satisfy users' needs.

The evaluators may take notes along the trip so as to remind themselves when giving feedback. The project team will somehow be

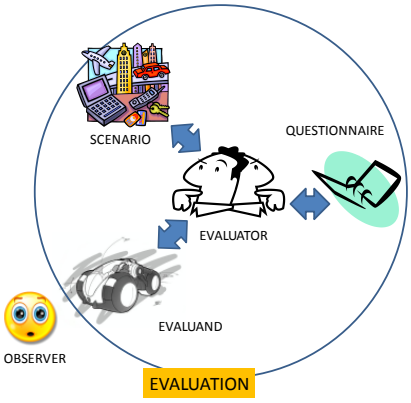

Figure 8 : CSG – Mobile - Evaluation

available supporting the evaluators.

The evaluators will fill the forms containing the questionnaires.

As reported, three different evaluations will be run in three different scenarios conducted by three different project partners: VOD, SMIND and TID.

4.3.3. Post-evaluation

In this stage the project team recovers an active role. The team analyses the input collected during the evaluation in order to draw useful conclusions and recommendations.

Figure 9 : CSG – Mobile – Post-evaluation

The questionnaires will allow two different ways of collecting feedback: questions to be answered with a number in a scale ("Likert questions"), and open questions to be answered with free text.

The first type of questions allows extracting graphic views of the value appreciated. The second type of question allows identifying issues or aspects that turn into recommendations.

Since there will be three different users group working over a common evaluand, a common questionnaire, the data comparison as well as data union will enable valuable conclusions.

The questionnaire will derive as a direct result of the assessment dimensions, since it is intended that these should be captured from the post-evaluation process. In Appendix A, a preliminary model of the questionnaire for the CSG trial participants is provided. The final questionnaire will be elaborated based on the final running release of the mobile application put to the test.

4.4. Additional Evaluations

WeKnowIt consortium has arranged two additional evaluations to be run by Vodafone in Athens and SMIND[1] in Krakow. Details are provided in next subsections. These two evaluations embrace all applications in the CSG scenario, i.e. pre-travel, travel and post-travel applications.

4.4.1. Athens evaluation trials

Participants of Athens' trials will be freely allowed to sightsee and explore places and landmarks of their own choice and at the same time follow the recommendation guidance of the CSG mobile application. As suggested by the pre-travel portal, the majority of most swarmed-in places of interest in Athens are congregated around its *historical centre* (Figure 10 and Figure 11). Therefore, for the case of those participants that will not schedule their sightseeing plan beforehand and through the guidance of the pre-travel portal, it will be hinted that they start their tour in the area of Athens' centre (Figure 10, Figure 11). This kind of handling would be favourable to the wanted collection of evaluation datasets and subsequently to the better assessment of the learning algorithms.

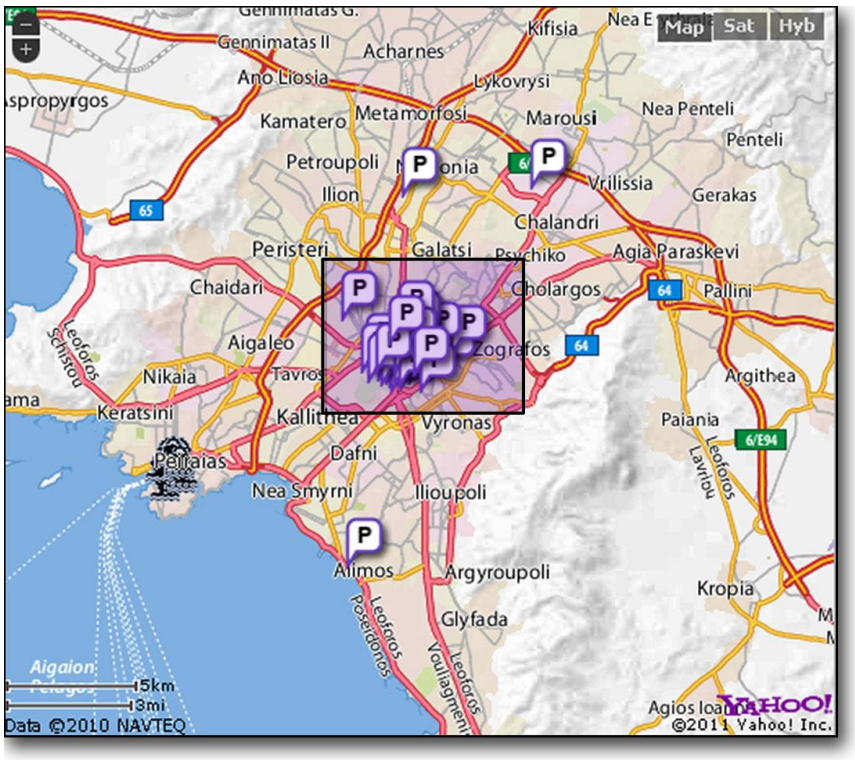

Figure 10: CSG - Pre-travel suggested POIs of Athens

[1] Project Partner – Software Mind SP. Z O.O (http://www.softwaremind.pl)

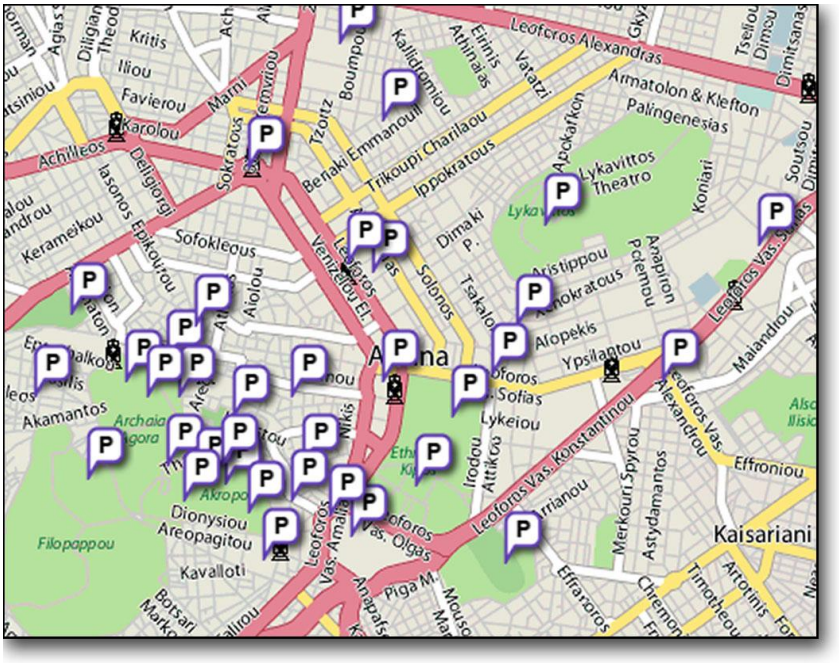

Figure 11: CSG - Close-up of the POI-crowded area (see rectangle of Figure 10); depicted size: 4 x 3.6 Km

Participants

Evaluation tests in Athens will involve conducting both internal tests as well as trials by an external group of testing consumers. For the internal tests at least 6 participants will be enrolled; similarly, 8 to 10 people, external to the company, will be enlisted as evaluators and provided with free data access to Vodafone's public network. Participants of both evaluation groups (i.e. internal and external) will be selected by the criterion of their mobile apps usage experience, which should be varying from that of an average/infrequent user to an experienced/ frequent user.

Equipment and network infrastructure

A combination of the supported handsets will be used for the Athens trials. These will include iPhones and smartphones running on Android OS, releases 2.1 and above.

Figure 12: CSG - App compatible smartphones

As far as network connectivity, the public mobile network of Vodafone Greece will be utilised. Vodafone commercially provides a state-of-the-art mobile network technology for data connectivity, offering a certified fastest mobile internet experience in the Greek telecoms market. Athens centre (including the trials area, cp. Figure 11) has a full coverage of either HS(D)PA or HSPA+ radio access technologies, that offer peak downlink speeds of 14.4Mbps and 28.8 Mbps, respectively, and uplink speeds of up to 5.8Mbps.

All participants will be given free data access to the public mobile network and for the specified duration of the trials.

4.4.2. Krakow evaluation trials

Kraków Evaluation Trail will be realised following the pattern given in Athen's trail description. Participants are expected to explore two major sightseeing districts of the city: Old Town and Kazimierz (see Figure 13). No strict route is planned; the participants are free to choose their path in the selected areas.

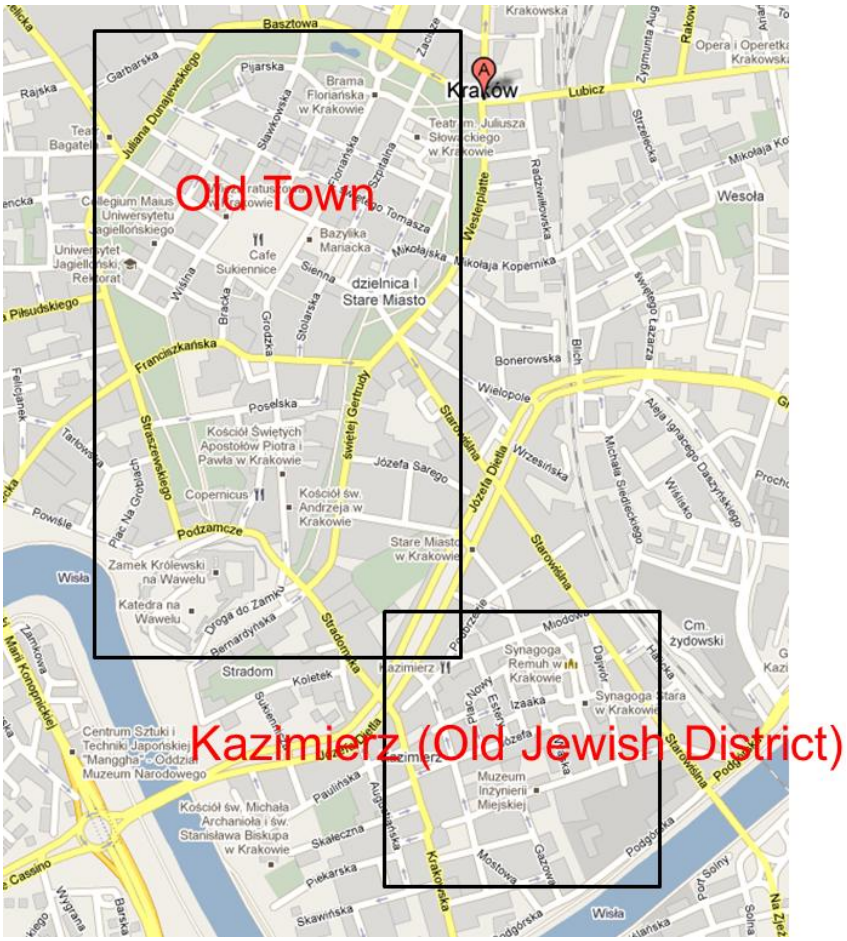

Figure 13: CSG - Kraków - Old Town and Kazimierz

Evaluation tests in Kraków will also be based on two groups of users:

- ~4 SMIND employees,
- ~6 people external to the company.

Kraków trail's participants will be selected using similar traits as described in Athens trail (i.e. mobile usage experience). All participants will be equipped with HTC Hero 2 smartphones.

5. Conclusions

The partners involved in the evaluations have cooperated fruitfully to produce this document describing the evaluation trials protocols to apply in the second round of the WeKnowIt project.

A methodological overview on evaluation methods serves as basis for the Emergency Response and Consumer Social Group protocols.

The protocol to be applied in the ER evaluation trials adopted a user-centred design methodology, where users were involved in every stage of the development process. To this end, user studies were started to collect early feedback – these studies included cognitive walkthrough, field trials, user interviews and focus group. The final evaluation will consist of a longitudinal study and evaluation of standard usability.

The protocol to be applied in the CSG evaluation trials doesn't follow a research approach but to know the user satisfaction, therefore the feedback is focused on usage dimensions. The pre-travel and post-travel tools will use standard usability questionnaires, and the mobility guidance tool will use ad-hoc questionnaires to explore dimensions such as availability, usability, etc in order to know how user needs and preferences are satisfied.

6. References

[1] WeKnowIt Annex I – "Description of Work", ver.1, FP7-215453, 26 Oct. 2007

[2] Bevan, N. What is the difference between the purpose of usability and user experience evaluation methods?

[3] Byström, K. Task complexity, information types and information sources. Doctoral Dissertation. Tampere: University of Tampere. (Acta Universitatis Tamperensis 688), 1999.

[4] Byström, K. & Järvelin, K. Task complexity affects information seeking and use. Information Processing and Management, 31(2), 191-213, 1995.

[5] Ciravegna, F. Dingli, A. Petrelli, D. and Wilks, Y.: User-System Cooperation in Document Annotation based on Information Extraction. in Asuncion Gomez-Perez, V. Richard Benjamins (eds.): Knowledge Engineering and Knowledge Management (Ontologies and the Semantic Web), Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02), 1-4 October 2002 - Siguenza (Spain), Lecture Notes in Artificial Intelligence 2473, Springer Verlag

[6] A. Dix, J. Finlay, G.D. Abowd and R. Beale, "Human-Computer Interaction", Prentice Hall, 1993.

[7] Jameson, A.; Spaulding, A.; and Yorke-Smith, N. 2009. Introduction to the Special Issue on Usable AI. AI Magazine 30(4).

[8] Kelley, J.F., "An empirical methodology for writing user-friendly natural language computer applications". Proceedings of ACM SIG-CHI '83 Human Factors in Computing systems (Boston, 12-15 December 1983), New York: ACM, pp. 193-196.

[9] Likert, Rensis "A Technique for the Measurement of Attitudes". Archives of Psychology 140, pp. 1–55, (1932).

[10] Mendoza, V., and Novick, D. G. Usability over time. In ACM 23rd International Conference on Computer Documentation, ACM Press (2005), 151-158.

[11] J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces", Proc. ACM CHI, Seattle, 1990, pp 249-256.

[12] Petrelli, D., Lanfranchi, V., Ciravegna, F.. Working Out a Common Task: Design and Evaluation of User-Intelligent System Collaboration, In Proceedings of Tenth **IFIP TC13** International Conference on Human-Computer Interaction (INTERACT 2005), Rome , 12-16 September 2005.

[13] Roto, V. User Experience Research in the Design and Development Phase. Keynote at the User Experience & User Generated Content workshop, Salzburg, Austria, 2008

[14] Vakkari, P., & Hakala, N. Changes in relevance criteria and problem stages in task performance. Journal of Documentation, 56(5), 540-562, 2000.

[15] Vaughan, M., & Courage, C. SIG: Capturing longitudinal usability: What really affects user performance over time? Extended abstracts on human factors in computing systems, 2149–2152, 2007..

[16] WeKnowIt "D7.1 Consumer and Emergency Response Use Case Initial Requirements", Sep. 2008

[17] WeKnowIt "D7.2 Emergency response and consumers' social group case study design and specification", Feb. 2009

[18] WeKnowIt "7.5.1 Consumer and emergency response use case first evaluation report", February 2010.

[19] WCAG 2.0 Guidelines, http://www.w3.org/TR/WCAG20/

A.1. CSG Mobile Guidance Questionnaire

weknowit

Evaluation Questionnaire – Mobile Guidance Application D7.5.2

Can you specify your equipment for the evaluation?

SIM card provider	Phone	Browser

The following items of functionality were **available**

Functionality Items	Yes	No	Sometimes	Not explored
Access				
Login				
Leaving				
Around you				
POIs				
Assess POI				
Show gallery				
Add image				
Mass Recommend related POIs				
Events				
Clusters				
Show gallery				
Group				
Chat				
Position				
Recommendations				
Personal				
Group				
Bookmarks				
Favourites				
Pre-travel				
Other				
Explore places				
Settings				

The mobile guidance application was **responsive**

agree	partially agree	neutral	partially disagree	disagree

- If there is no level of agreement, would you please give details on the lack of responsiveness? (Did you check network coverage?) – Can you point out non-responsive functionality items?

1

weknowit

Evaluation Questionnaire – Mobile Guidance Application D7.5.2

The mobile guidance application was easy to use

agree	partially agree	neutral	partially disagree	disagree

- If there is no level of agreement, would you please inform what areas or elements were difficult to use?

The mobile guidance application was easy to understand

agree	partially agree	neutral	partially disagree	disagree

- If there is no level of agreement, would you please inform what items, behaviours, information or elements were confusing?

The appearance of the mobile guidance application is appealing

agree	partially agree	neutral	partially disagree	disagree

- If there is no level of agreement, would you please make suggestions or spot what areas or elements deserving improvements?

2

weknowit

Evaluation Questionnaire – Mobile Guidance Application D7.5.2

The following items of functionality were **useful**

Functionality Items	Yes	No	Sometimes	Not explored
Access				
Login				
Leaving				
Around you				
POIs				
Assess POI				
Show gallery				
Add image				
Mass Recommend related POIs				
Events				
Clusters				
Show gallery				
Group				
Chat				
Position				
Recommendations				
Personal				
Group				
Bookmarks				
Favourites				
Pre-travel				
Other				
Explore places				
Settings				

All items of functionality are **suitable** for mobile guidance

agree	partially agree	neutral	partially disagree	disagree

- If there is no level of agreement, would you please identify what were not suitable?

3

weknowit

Evaluation Questionnaire – Mobile Guidance Application D7.5.2

You have missed other features or functionality items suitable for mobile guidance

agree	partially agree	neutral	partially disagree	disagree

- If there is any level of agreement, would you please make suggestions?

What are the most interesting features or functionalities you found while using the mobile guidance application?

Would you rate your overall usage experience?

very positive	partially positive	neutral	partially negative	very negative

- Would you point out those aspects or items influencing positively and those aspects or items influencing negatively?

We would appreciate any other additional comment on the evaluation process or the mobile guidance application itself.

4