



WeKnowIt

Emerging, Collective Intelligence for Personal,
Organisational and Social Use

FP7-215453

D7.6.1

Initial consumer and emergency response use case evaluation protocols

Dissemination level	Public
Contractual date of delivery	Month 18, 30-09-09
Actual date of delivery	14-10-09
Workpackage	WP7, Case studies
Task	T T7.1.3, T7.2.3, Evaluation and user
Type	Report
Approval Status	Draft
Version	0.15
Number of pages	57
Filename	D7.6.1_2009-10-14_v15_TID
Abstract:	<p>This deliverable describes both consumer social group and emergency response case evaluation protocols what will be used for the evaluation of the first WKI prototype.</p> <p>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.</p>



co-funded by the European Union

History

Version	Date	Reason	Revised by
0.1	20/07/2009	TOC	TID
0.2	18/08/2009	First Pass at ER and gen. Methodology	USFD
0.21	21/08/09	Amendments to ER and Annex	SCC
0.3	31/08/09	First contribution at CSG and general. 1st integrated version	TID
0.4	02/09/09	Second contribution at CSG. Amendments to CSG Questionnaires and Methodology.	TID
0.5	10/09/09	Second contribution by USFD	USFD
0.5.1	12/09/09	Revision and Amendments at CSG and general.	YAHOO
0.5.2	13/09/09	YAHOO and USFD contributions have been integrated. TID 's revision at CSG. 2st integrated version	TID
0.5.3	13/09/09	Updated ER assessment - clarified overview process.	USFD
0.6	15/09/09	Added the Introduction and the conclusions. Updated CSG-Sub-Scenario Mobile guidance assessment. 3rd integrated version: Deliverable ready for being internal reviewed.	TID
0.7	21/09/09	Internal review	Tomasz Kaczanowski, (SMIND)
0.7b.	22/09/09	Internal review.	Costis Contopoulos, (VOD)
0.8	23/09/09	TID 's revision at CSG and general according to SMID and VOD comments.	TID
0.9	24/09/09	USFD revision to address reviewers comments	USDF
0.10	27/09/09	TID revision to address reviewers comments	TID
0.10b	28/09/09	Modification on the CSG evaluation Workplan.	TID
0.11	29/09/09	Deliverable ready for getting WP 's contribution about the protocols to test the WeKnowIt collective Intelligence	TID
0.12	06/10/09	First draft of altered version of ER evaluation protocol	USDF
0.13	09/10/09	Updated WeKnowIt 's general assessment issues to include the protocols to test the collective Intelligence	TID

Version	Date	Reason	Revised by
01.4	12/10/09	Updated general assessment and the ER assessment	USDF
0.15	14/10/09	Final deliverable ready for being submitted.	TID

Author list

Organization	Name	Contact Information	Reason
TID	Ana Flores Cuadrado	anafc@tid.es	Author
USFD	Simon Tucker	s.tucker@dcs.shef.ac.uk	Author
SCC	Sheila Slingsby	Sheila.slingsby@sheffield.gov.uk	Author
YAHOO	Roelof van Zwol	roelof@yahoo-inc.com	Author
SMID	Tomasz Kaczanowski	tomasz.kaczanowski@softw.aremind.pl	Internal reviewer
VOD	Costis Kontopoulos,	Costis.Kontopoulos@vodafone.com	Internal reviewer

Timeplan before delivery

Next Action	Deadline	Care of
Agreement on the TOC; statement of partners' interest and allocation of work	24/07/2009	all
1 st contributions received	21/08/2009	All
1 st integrated version	31/08/2009	TID
2 nd contributions received	2/09/2009	TID
3 st contributions received	12/09/2009	All
2 nd integrated version	13/09/2009	TID
Any other issues before final contribution	14/09/2009	All
Deliverable ready for being internal reviewed.	16/09/2009	TID
Reviewer contributions received	22/09/2009	SMID,VOD
Revision to address reviewers comments	28/09/2009	ALL
WP 's contribution about the protocols to test the Weknowit collective Intelligence	09/10/2009	ALL
Final deliverable ready for being reviewed and submitted.	14/10/2009	TID

Executive Summary

The purpose of this deliverable is the description of consumer social group and emergency response evaluation protocols what will be used for the evaluation of the first WKI prototype. To achieve this goal, first, we have outlined the methodology that will be used to evaluate the WKI system. The **methodology** has been chosen like the continuation of the user-centered methodology followed in the tasks T7.1.1 and T7.2.1. In this methodology, the mini-scenarios, a subset of the full functionality of WeKnowIt of each scenario, will be used to identify the procedures for evaluation that are needed to test the system and the results expected, and to identify the user groups that interact with the system.

For the **Emergency Response Case**, two different evaluations will be used: One for Sheffield citizens, and one for ER Staff. In the context of **Consumer Group Case**, we have considered two different types groups of users for the evaluations: One represented by the users using the desktop device and the other one by the users using a mobile device. However, for the evaluation of the CSG prototype, preferentially the same group of subjects should be used across the three phases of the travel scenario

The correct functionality of first prototype will be evaluated by performing functional tests and getting the user's feedback with specific evaluation questionnaires, and interviews, as well. Besides, questionnaires, video recording and informal interviews will also be used to determine the user reaction to the system and to know if E-WKI met the required success measures.

Finally, the first prototype will have a subset of the proposed full functionality of the final system, therefore, the protocols for the prototype evaluation described in this document are only a preliminary version, they will be extended in the final deliverable D7.6.2 (M33).

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

1. Introduction	9
1.1. Goal and Scope	9
1.2. Structure of the Document	10
2. Methodology	12
2.1. Non-Participatory Evaluations	12
2.1.1. Functional Evaluation	12
2.1.2. Cognitive Walkthrough	12
2.1.3. Heuristic Evaluation	12
2.2. Participatory Evaluation	13
2.2.1. Field Trials	13
2.2.2. Lab Studies	14
2.3. Measurement Techniques	14
2.3.1. Think Aloud	14
2.3.2. Video Recording	14
2.3.3. Logging	15
2.3.4. Eye Tracking	15
2.3.5. Screen Recording	15
2.3.6. Physiological Recording	15
2.4. Post Hoc Measurement Techniques	15
2.4.1. Questionnaires	16
2.4.2. Interviews	16
2.5. Evaluation methodology for the WKI System	16
2.5.1. Emergency response Case	16
2.5.2. Consumer Group Case	17
3. Emergency Response Case Study	19
3.1. Scenario Definition	19
3.2. Scenario Constraints and Requirements	22
3.2.1. Technological Constraints	22
3.2.2. Temporal Constraints	22
3.3. Target user groups	22

3.3.1.	Emergency Response Professionals	23
3.3.2.	Sheffield citizens.....	23
3.4.	Evaluation	23
3.4.1.	System Checks	25
3.4.2.	Recruitment of Participants	25
3.4.3.	Objective Data Collection	25
3.4.4.	Sheffield Citizen Subjective Evaluation	26
3.4.4.1.	Subjective Evaluation Tasks	26
3.4.4.2.	Success Measures	26
3.4.5.	Emergency Response Staff Evaluation	28
3.4.5.1.	Subjective Evaluation Tasks	28
3.4.5.2.	Task Measurements.....	28
3.4.6.	Citizen Objective Evaluation	29
3.4.7.	ER Professional Objective Evaluation	30
3.4.8.	Evaluation Report	31
4.	Consumer Social Group Scenario	32
4.1.	Scenario Definition	32
4.1.1.	Sub-Scenario Travel preparation.....	32
4.1.2.	Sub-Scenario Mobile guidance	33
4.1.3.	Sub-Scenario Post travel experience	33
4.2.	Evaluation	34
4.2.1.	Sub-Scenario Travel preparation.....	34
4.2.1.1.	Evaluation Tasks	34
4.2.1.2.	Success Measures	34
4.2.2.	Sub-Scenario Mobile Guidance Stage.....	34
4.2.2.1.	Evaluation Tasks	34
4.2.2.2.	Success Measures	35
4.2.3.	Sub-Scenario Post travel experience	36
4.2.3.1.	Evaluation Tasks	36
4.2.3.2.	Success Measures	36
4.3.	Scenario Constraints and Requirements	36
4.3.1.	Technological Constraints	36
4.3.2.	Temporal Constraints	36

4.4. Target user groups 37

4.5. Evaluation Work Plan 37

5. Conclusions 40

6. References..... 41

A. Overall questionnaire for both scenarios 42

B. Emergency Response Case 43

 B.1. Subjective Task Sheet for Sheffield Citizens..... 43

 B.2. Questionnaire for Sheffield Citizens (Subjective Evaluation)..... 44

 B.3. Task Sheet for ER Staff (Subjective Evaluation) 44

C. Consumer Social Group Questionnaires 57

List of Figures

Figure 1: Schematic for 1st prototype evaluation plan 24

Figure 2: Evaluation Work Plan for the CSG evaluation 37

Figure 3: CSG Evaluation Architecture 39

1. Introduction

1.1. Goal and Scope

The main objective of WeKnowIt 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 has chosen two radically different but complementary use cases to demonstrate the wide applicability of its technologies and research activities: Emergency Response case, Consumers Social Group case study. After the current situation was analysed, a phase of requirements gathering proceeded, followed on by the writing of the scenarios for each of the two use cases [6] [7]:

- In particular, in the **Emergency Response** (ER) scenario *citizens* distributed across the region will be able to participate in the monitoring of an (emergency) incident or event. This will benefit *Emergency Response planners* that will have real time information available on which they can base their decisions and strategies, enabling them to better react to an emergency. Moreover, the system will automatically gather information available elsewhere on the network to aid the Emergency Response, thus making possible for an emergency planner to find exactly the required knowledge amongst all the available information and to selectively make this knowledge available to the citizens (e.g. information about open roads, information about relatives involved, etc.) in a largely automated way. The technologies in use will therefore also encourage and enable dialogue between the Emergency Responders and individuals, groups and communities
- For the **Consumer Social Group** (CSG) case study WeKnowIt will provide the ability to extract meta-information from various content sources and user generated content within the system and will furthermore provide enhanced publishing tools to support travel activities for *individuals and groups*. WeKnowIt will automatically analyze the uploaded content, user's profiles and their actions in order to improve knowledge about the community. Based on the travel behaviour of all users and an analysis of their profiles, WeKnowIt will extract information about the most preferred travel destinations and the trends and changes in user behaviour. This information will be used to make recommendations to the users and support the decision making process before the trip (e.g. choice of destination planning) and during the trip (e.g. going to an on-site restaurant, etc.).

The purpose of this deliverable is the description of consumer social group and emergency response evaluation protocols what will be used for the evaluation of the first WKI prototype. The objectives of the evaluation task and the evaluation protocols are the following:

- An evaluation of the current status of WKI first prototype.
- An evaluation of the benefit of the Collective Intelligence processing implemented for the first prototypes of WeKnowIt.

For the evaluation of the status of E-WKI's first prototype, tests will be performed checking the system works properly, testing how well the integration between of WKI services and user interface works according to the architecture proposed. This will test whether the concepts around the case studies are valid and whether the system has met these requirements at the user and system level.

The secondary aim of the evaluations is to objectively and subjectively assess the benefits that Collective Intelligence brings to the two use cases outlined in WeKnowIt. This will take the form of a comparison between the users' ability to make decisions and make assessments on the basis of the information when presented by WeKnowIt and when presented to the users in a raw form. The purpose of this test is to measure the effectiveness of Collective Intelligence, as implemented by WeKnowIt.

1.2. Structure of the Document

The deliverable has been organized in the following way: After presenting the deliverable overall approach, the **second chapter** presents different approaches to evaluating interactive systems, focusing specifically on WeKnowIt system and outlining the methodology that will be used for its evaluation. The **methodology** has been chosen like the continuation of the user-centered methodology followed in the tasks T7.1.1 and T7.2.1 (refer to WP7 deliverables [6] [7]). In this methodology, mini-scenarios will be used to identify the procedures for evaluation that are needed to test the system and the results expected, and to identify the user groups that interact with the system. Questionnaires, video recording and informal interviews will also be used to determine the user reaction to the system and to know if E-WKI met the required success measures.

In the **third and the fourth chapter**, we present the protocols for the evaluation of both studied cases (ER and CSG). While the Emergency Response protocols are presented in **Chapter 3**, **Chapter 4** is dedicated to the definition of the Consumer Social Group protocols. In both cases we provide high-level scenario overviews and set the background on the scenario definition approach, the outline of the constraints and the requirements and the identification of the target user groups for the evaluation specified in the following sections. Finally, we conclude by defining the protocols for the evaluation of every case.

Chapter 5 gives a briefly conclusion and provides an outlook on the further work to be carried out in the subsequent stages of the current work package; finally the questionnaires for the user’s interviews are presented in the **Annex**. Annex A provides more general usability questions concerning both cases, while Annexes B and C contain the specific questionnaires that refer to each use case, in particular.

The evaluation protocols described in this document are an initial approach to assessing the first version of prototypes and during their first trial phase. The results of this evaluation will be extended with the ensuing assessment – to be delivered in D7.6.2, by month 33 – by which the final version of prototypes is to be validated.

2. Methodology

This section begins by describing different approaches to evaluating interactive systems [2], continues by describing the options for participatory evaluation and concludes by outlining the methodology that will be used to evaluate the WKI system. 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 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.

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 mockups 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 [3]. 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.

2.2.2. 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 [3] 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.

2.3. 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 [4] 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.

2.5. Evaluation methodology for the WKI System

The first prototype of E-WKI will have a subset of the proposed full functionality of E-WKI, as outlined in the mini-scenarios presented in the following sections. Given the prototypical nature of the prototype an initial lab study will be used to ensure that the system is functioning correctly and that the users are both able to interact with the system and understand what the system is asking of them at each stage. This lab study will be used to validate that the system is able to meet the requirements laid out in the mini-scenario.

Following the lab study lightweight comparative studies will be additionally be used. These field trials will focus on whether WKI is able to process field data and present this data effectively to the end user. These trials will additionally address the question of whether the Collective Intelligence implemented by the WKI system adds value to the information handling of the user groups that interact with the system.

2.5.1. Emergency response Case

The Emergency Response Case will be evaluated using equipment and personnel from the University of Sheffield (USFD). Preliminary lab based

evaluations will be used to ensure that the WKI system is functioning correctly. These will take the form of purely functional tests to ensure that the system code is robust. The final step of the preliminary tests is to use a standard set of web usability heuristics [8] in order to ensure that the WKI interface meets these standard access requirements and that the system functions according to the mini-scenario specification. Following these preliminary evaluations a more formal evaluation of the system will take place.

The mini-scenario specification will be used to identify the user groups that interact with the system and what tasks they perform. The evaluation will then use small groups of users that are representative of these target user groups and have them carry out the tasks specified in the mini scenario. This subjective stage of evaluation will be used to ensure that the users of E-WKI are able to use the system effectively. Questionnaires and interviews will be used to assess how effective E-WKI was in assisting the users in carrying out these tasks.

A secondary evaluation will be then be used to assess how the Collective Intelligence processing enhances both user groups understanding of emergency situations. This process will be carried out with both sets of users: In the Citizen case the question will be how Collective Intelligence contributes to the citizen's understanding of a single emergency, for ER professionals the question will be how Collective Intelligence contributes to the professional's interpretation of information relating to multiple emergencies taking place in a typical city.

2.5.2. Consumer Group Case

The methodology followed for the **Consumer Group Case Scenario** evaluation will be similar to the one followed in the ER case. The user groups that interact with the system and the tasks they perform will be identified by examining the mini-scenarios. In the context of Consumer Group Case scenario, there is no difference between the users of the system. The only difference between the users is the type of device (mobile or desktop PC) they are using to test the scenario and their location, that is the way, we have preferred to split the evaluation task into several different parts.

Additionally, the correct functionality of the first prototype will be evaluated by means of the development of some functional tests. In this case, the previous mini-scenario analysis will also help us to identify the **cognitive walkthroughs** that are needed to test the system and the results expected.

On the other hand, and as far as the CSG is concerned, evaluation questionnaires will be used to determine the user reaction to the system and the interface. They will also be used to determine whether the WKI system has met the required success measures. We will use the SUS (**Usability & Satisfaction**) questionnaires included in the Annex C, which

contains specific questions about the CSG functionality and concerning the content that users are going to evaluate. We will use these questionnaires to get the user's feedback about the specific functionalities.

It is also possible to make other types of evaluation, such as tests or benchmarks to check status of some system performance variables (for example the response rate). However, we think the realization of these tests will be more suitable for the evaluation of the whole CSG functionality in the next iterations, when the GSG functionality is fully developed and WeKnowIt's different intelligence layers have been integrated properly.

Finally, to test the effectiveness, and efficiency of the CSG prototype the test subjects can be presented with a series of tasks, and the degree of success in executing these tasks within a predefined time window can be measured, using logging facilities of the prototype to monitor the users' behavior.

Preferentially, the evaluation of the CSG prototype will use the same group of subjects across the three phases of the travel scenario. Initially, the user will do some "research" prior to the visit. The same subject then goes on location and finally during the trip or after returning back home they can participate in the post-travel experience.

3. Emergency Response Case Study

The Emergency Response (ER) scenario concerns itself with how E-WKI should respond to a flooding incident in a city. Below, the scenario is defined, the constraints and the requirements that the scenario makes are outlined, the target user groups are identified and an evaluation protocol for this case study is specified. The mini scenario outlined below is formed as a functional subset of the main scenario defined in D7.1. The purpose of the mini-scenario is to outline a minimal set of interactions which are required to demonstrate and evaluate that the Collective Intelligence implemented by WKI is sound in terms of methodology and usability.

Whilst the subjective evaluations will be centred on the flooding scenario, alternative emergency situations will be used in the objective evaluations in order to ensure that the WKI system is evaluated on a range of data sets and emergency scenarios.

3.1. Scenario Definition

The scenario is split into a number of discrete events and describes the events that occur during the scenario. Note that not every event has a corresponding action for the WKI system and some events may result in multiple interactions. For full details of this scenario and for clarification about the technical terms, see D7.1

1. Day 1, 15:00

A call from the Fire Service is received by the Emergency Response office in the Sheffield City Council, warning about a potential flood emergency.

2.

John, an experienced member of the ER staff, is taking the call and, as he thinks it may evolve to be a major emergency, he puts resources on standby.

3.

After a few minutes John receives two more calls, one by the Police Chief and one by the Ambulance Services, both describing the floods. A major emergency is declared and John decides to activate plans and resources.

4.

John decides to activate the team and sets up a “Gold Team” to strategically manage the emergency:

- Establish a gold strategy for managing the event
- Communication with the media
- Communication with other agencies
- Prioritise requests
- Prepare a long-term recovery strategy

5.

John also activates other command structures at both Silver and Bronze level. The Silver level is the SCC Emergency Control Centre and is the tactical operational command structure that will manage the events and Bronze level control is situated at the incident site(s).

6.

John will be part of the Silver tactical team. Lucy will be responsible for the Gold team. Andrea is sent by Lucy to the disaster scene as part of the Bronze Team.

7. Day 1, 17:00

Mark, a Sheffield citizen, is still at work and getting ready to go home. When exiting the office, he realises the road is flooded and it may be very hard to reach the parking.

8.

Mark calls the emergency service, is asked to describe the situation.

9.

The emergency service asks Mark, whether his mobile phone is equipped with a digital still camera and whether he can take a picture of the scene and upload it to E-WKI. While Mark is talking the emergency service inputs his name into the "New user" form in the system. A message is sent to Mark with the invitation to E-WKI containing login details.

10.

After Mark took a picture of the road, he uploads the photo using the web interface. Mark also quickly tags the picture by entering the word "flood".

11.

Immediately the image is analysed by the intelligent system and categorised as "flooding".

12.

As someone else has already added a picture tagged as flooding from a nearby geographic location the system automatically clusters them together.

13. Day 1, 17:30

When John logs in (from his desktop PC) he gets a screen with an overview of the available information.

14.

When opening the main page John views the information using a geographical visualisation. John can see the location of the uploaded photos. This helps John in having a better overview of the situation.

15.

John sends Andrea (a Forward Liaison Officer) to the scene of emergency to take pictures.

16.

When Andrea arrives on the scene of the emergency she starts taking pictures and uploads them to E-WKI.

17. Day 1, 18:00

John starts thinking about making some of the information public so normal users of the Sheffield City Council website can see the current situation and know which the most affected areas are. He then approves some of the images for public visibility

18.

John asks for an analysis and an overview of what was happening in a specified period of the incident. The metadata on the phone calls received by the city council as well as their content are taken into account.

19. Day 2

Mark is at home and wants to review all the information that he captured during the emergency and submit new additional photos that he did not submit yesterday.

20.

Mark has also been sent by friends some pictures about the emergency and wants to add them. Mark logs into E-WKI using his desktop PC. He has already a login to the site so the system immediately presents him his home page.

21.

Mark chooses to upload the new content. When the content is uploaded Mark decides to tag all his photos with the words "flood", "river", "road", "water". He also adds a short comment of how he got involved in the floods and what he saw.

22. Day 3

In his office John is dealing with the recovery phase. He still has to monitor the incidents but also start thinking to how the whole process can be managed. John receives a notification via mail that new content has been added to E-WKI so he logs in and browses the new content.

3.2. Scenario Constraints and Requirements

Here the technological and temporal constraints and requirements placed on the evaluation are discussed.

3.2.1. Technological Constraints

E-WKI is accessed over both mobile devices and from desktop PCs. Whilst web access from desktop PCs is largely homogeneous (since the means of accessing the web across different equipment is largely similar), different classes of mobile devices exhibit significant differences in terms of display space, the rendered layout of formatted web content and the method of interaction. In evaluating E-WKI the system should be run over different devices as well as different types of users. To effectively test the performance of the mobile interface, users who already have suitable mobile devices for the first phase of evaluations will be explicitly recruited.

Because of the homogeneity of desktop PCs it will be sufficient to perform a cognitive walkthrough on the four main browser types (Firefox, Safari, Internet Explorer & Opera) in order to ensure that the system functions correctly across these browsers.

3.2.2. Temporal Constraints

The scenario defined above occurs over different times and places. There is, therefore, a question about whether the evaluation needs to reflect these differences. Clearly, since this evaluation will take the form of a lab study, there is no requirement to replicate space when evaluating E-WKI. In addition, since the temporal changes are not part of the interaction with the system it is unnecessary to include corresponding temporal breaks when evaluating E-WKI.

In the comparative evaluations, for practical reasons, the data collection phase will take place at a different time to the evaluation phase. This will ensure that a representative set of data can be collected and will mean that the evaluation is fair in its assessment of the impact of CI on the information processing of E-WKI users.

3.3. Target user groups

The following individuals were involved in the mini-scenario presented above:

ER Staff:

- **John:** Member of the Silver Team, an experienced member of the ER staff.
- **Lucy:** Member of the Gold Team (*does not interact with E-WKI*).
- **Andrea:** Member of the Bronze Team.
- **Mark:** A Sheffield citizen.

- **Emergency Service Call Handler:** handles Mark's emergency call.

Given that E-WKI will not integrate with the emergency services system and that the task performed by the call handler (essentially creating an account for Mark) will be handled relatively automatically, it will not be necessary to explicitly evaluate this step of the process. Therefore for this evaluation there are two distinct groups of users.

3.3.1. Emergency Response Professionals

Since the experience and knowledge of ER professionals will have a significant bearing on how the WKI system should function and how information should be presented. Therefore the evaluation of any part of the system that ER professionals interact with should be carried out by ER staff. Since the group of users is of a limited size the approach taken with ER professionals will focus on subjective data in the form of interview in order to maximize the information collected by the evaluations.

Note that the term ER staff applies to any emergency response organization in the UK, not just members of the SCC but members of the fire and police services whose job requires them to make response decisions on the basis of incoming information.

3.3.2. Sheffield citizens

The group of users represented by Sheffield citizens can be limited to citizens who have internet access and a suitable mobile device and therefore this group can be recruited from technologically aware members of the general public. Since the University of Sheffield is carrying out the main evaluations, Sheffield based internet forums should be used to ensure the recruitment is not biased towards students and staff of the University.

The selection of participants will ensure that people with a range of ages will participate in the evaluations so that the system is tested with older users as well as younger. Users will be encouraged to carry out the evaluation on their personal mobile devices wherever possible.

For purposes of the recruitment of participants we will also consider evaluating the system with volunteers who would be willing to participate from amongst partner organizations, local businesses, Emergency Planning Society, local churches, Major Incident Response Group workers, voluntary partner groups. This will enable the system to be evaluated with a wider range of participants and will also act as a demonstrator of the system with such groups.

3.4. Evaluation

The schematic for the evaluation plan is shown below.

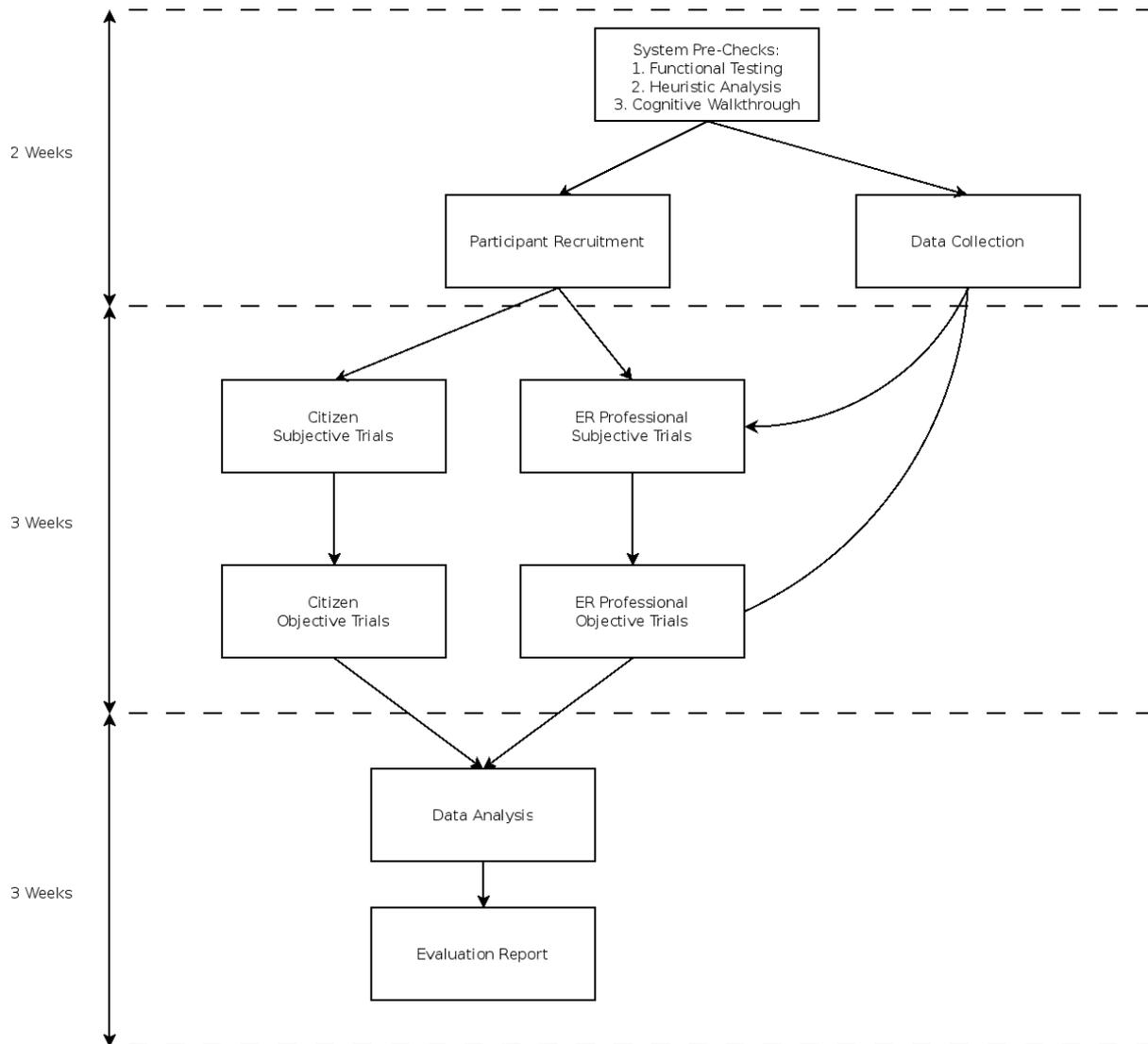


Figure 1: Schematic for 1st prototype evaluation plan

The evaluation begins with some system checks which are used to ensure that the overall system functions as intended and that the subsequent trials can commence. At this point a short data collection phase is used to collect some field data that can then be used in the evaluations.

For citizens there are two primary evaluations – a subjective test which ensures that the citizens are able to make use of the upload functionality and that as typical users of E-WKI they are able to understand the functionality and interface of the system. There is then a further test which is designed to measure how well the Collective Intelligence present in WKI supports the post-event information processing by the citizen.

For the ER Professionals a similar process occurs. There is firstly a subjective evaluation which determines if the participant is able to use the system correctly and perform the tasks specified in the mini-scenario. There is then a similar objective evaluation which is used to investigate how the CI processing supports the ability of the ER Professional in understanding three events occurring simultaneously in a typical city.

These stages are described in detail below.

3.4.1. System Checks

The system checks are formed from four different assessments. Firstly throughout the development of E-WKI functional testing will be used to ensure that the system functions at a low level correctly. Following the completion of the system heuristic evaluations and cognitive walkthroughs will be used to ensure that the system functions correctly with regards to the tasks outlined below. Specifically both Nielsen's Heuristics and the W3C standard will be used to ensure that the system is usable and meets standard interface guidelines.

3.4.2. Recruitment of Participants

Once the system checks are complete, participants will be recruited. For citizens there are two stages to the evaluation – 10 participants will be recruited for the first stage and 20 for the second stage. In the first stage participants will be running through the upload process in order to ensure that the process is understandable to the users, that they are correctly encouraged to tag and comment on their data and that the system correctly functions with multiple mobile devices. Thus the initial 10 participants will be chosen on the basis of owning an internet enabled mobile device. The second stage will focus on the citizen's recovery of information and will objectively evaluate the effect that CI has on their perception of emergency events in the city.

For the ER professionals a minimum of 4 participants will be needed. Because of the limitations of time and availability of ER professionals a smaller number will be used and so the evaluations will focus on interviews as a primary means of collecting information from this group of users. In addition, for the ER professionals, the same group of users will be used for both experiments.

3.4.3. Objective Data Collection

To adequately evaluate Collective Intelligence there is a requirement to have a suitable set of data that a) fits into the general WKI framework and b) simulates a realistic event somewhere within a city.

To facilitate this, a short data collection phase will be employed in the evaluation of the first prototype of E-WKI. Three sets of data will be collected, each containing 30 images and associated metadata concerning three events of differing severity which are occurring in different locations in the city. The events will not be simulated physically but the actors recording information about the events will be given instructions about the events, their role in them and they will be instructed to act as if the event is genuinely occurring. Specifically they will be instructed:

- From which direction to approach the event from. This is to ensure that the data from different actors reflect different perspectives of the emergency
- What kind of event it is and where it is occurring.
- What time the event is occurring.

They will then be asked to reach their approach point at the appropriate time and use the WKI upload interface to take five images of the event, to tag the images and submit them to E-WKI. They will additionally be asked to make a short comment describing the event.

Since the upload procedure will be evaluated elsewhere, this data collection need not be carried out by typical users of the system but if the subjective evaluations indicate that typical users differ significantly from the ideal (e.g. if typical users do not tag or comment on images) then the data should be suitably altered to reflect how typical users interact with the system.

In addition to these primary sets of data a dummy set will also be collected for use in training and the subjective evaluations presented below. The precise details of the data collection events and how they are to be presented to the users are given in the appendix.

3.4.4. Sheffield Citizen Subjective Evaluation

The purpose of the subjective evaluations is to determine that the WKI system is usable by typical users.

3.4.4.1. Subjective Evaluation Tasks

The interaction with E-WKI can be broken down into a number of discrete tasks (with reference to the steps outlined in the mini-scenario):

Using a mobile device:

- Step 9: *Log into the E-WKI System.*

An account is created for the citizen and using information contained in an email they log into E-WKI.

- Step 10: *Uploads an image.*

Assuming that the citizen has already taken a suitable image, the citizen uses E-WKI to upload the image to the system.

- Step 10: *Tags an image.*

Upon uploading the image the citizen is able to apply the tag “Flood” to the image.

3.4.4.2. Success Measures

From the above evaluation tasks measures of success can be generated. Note that for actions using the mobile device, the user will not have had any training with the system and so must be able to carry out the tasks

required of them without any assistance. This criterion is reduced in the desktop case since a help function could be used. The pre-test cognitive walkthrough results are also listed here.

For the purposes of this evaluation a dummy set of data will be used. Since this evaluation process is focused on the usability of the system there is no need for the system to display meaningful data.

Log into the E-WKI System. (Mobile)

The cognitive walkthrough should confirm that it is possible to log in to E-WKI with the required credentials.

- **Success:** The user is able to access E-WKI first time without assistance.
- **Partial Success:** The user is able to access E-WKI without assistance.
- **Partial Failure:** The user is able to access E-WKI with assistance.
- **Complete Failure:** The user is unable to access E-WKI.

Upload an Image. (Mobile)

The cognitive walkthrough should confirm that it is possible to upload an image.

- **Success:** The user is able to upload an image to E-WKI without assistance.
- **Partial Failure:** The user is able to upload an image to E-WKI with assistance.
- **Complete Failure:** The user is unable to upload an image to E-WKI.

Tag an Image. (Mobile)

The cognitive walkthrough should confirm that it is possible to tag an image.

- **Success:** The user is able to tag the image without assistance.
- **Partial Failure:** The user is able to tag the image with assistance.
- **Partial Failure:** The user does not tag the image.
- **Complete Failure:** The user is unable to tag the image.

Following the completion of these tasks an informal interview will address any problems that occurred during the evaluation. The purpose of this interview is to identify where the problems lay if the user was unable to complete some of the tasks. Following this interview the questionnaire should be filled out and another informal interview should follow up on any relevant answers to the questionnaire questions.

3.4.5. Emergency Response Staff Evaluation

3.4.5.1. Subjective Evaluation Tasks

As with the citizens, the E-WKI interaction is broken down into a number of discrete tasks:

Using a desktop device:

- Step 13: *Logs into the E-WKI system*
The ER staff should know their credentials for logging into the system.
- Step 13: *Collects an overview of the available information*
The ER staff member is given an overview of all the available information - this includes information from multiple events.
- Step 13: *Specific overview*
The ER staff member should be able to focus specifically on the "flooding" event
- Step 17: *Approves images for public viewing*
The ER staff member then selects some of images and changes their credentials so that they can be viewed by the public.

3.4.5.2. Task Measurements

Again, considering that the ER Staff will have had some training with the system, success criteria can be defined for each task. Since this approach is used to determine how well the system functions for specific information the extra set of data collected above will be used to carry out the evaluation.

Log into the E-WKI system

The cognitive walkthrough should confirm that it is possible to log into the E-WKI system.

- **Success:** The user is able to access E-WKI without intervention.
- **Partial Failure:** The user is able to access E-WKI with intervention.
- **Complete Failure:** The user is unable to access E-WKI.

Get an overview of the available information

The cognitive walkthrough should confirm that the system displays the overview upon login.

- **Success:** The ER Staff is able to interpret the overview of information presented to them.
- **Complete Failure:** The information displayed to the user constitutes a poor overview to the ER staff.

Get a specific overview of a single event

The cognitive walkthrough should confirm that it is possible to navigate to the flooding event view from the overview and the relevant information is displayed.

- **Success:** The information displayed to the user constitutes a good overview to the ER staff.
- **Partial Success:** The information displayed to the user constitutes a fair overview to the ER staff.
- **Complete Failure:** The information displayed to the user constitutes a poor overview to the ER staff.

Approve images for public viewing

The cognitive walkthrough should confirm that it is possible to alter the permissions of a number of images.

- **Success:** The ER staff member is able to approve a number of images for public viewing without intervention.
- **Partial Success:** The ER staff member is able to approve a number of images for public viewing with some intervention.
- **Failure:** The ER staff member is unable to approve any images for public viewing.

As with the Sheffield citizens, interviews will be used to address any problems that the user had with the interface and a questionnaire and follow up questions will be used to determine the Staff member's reaction to the interface and system.

3.4.6. Citizen Objective Evaluation

This part of the evaluation focuses specifically on how E-WKI presents information regarding an event and how well the citizen is able to analyse and process this information. By doing so they will form an overview of the event and will, therefore, be able to answer general questions about it. The objective evaluation will compare the understanding the citizen has of the event using E-WKI to that of understanding the event with raw data.

The evaluation design is a within subjects using the three data sets that were collected. The user will be told that they witnessed an emergency earlier in the day from a long distance and are now trying to find out what happened. They will be provided with one of the interfaces and will be asked to answer three questions about the event:

- Where the event took place
- When the event took place
- How severe the event was

In the *baseline* case the user will be given the raw comments that were made by the actors. In the *low informational* case the user will see the full

set of data that was uploaded to E-WKI (i.e. the images, tags and comments) in paper form. This simple interface simulates the information processing currently required by the citizen – for example searching the web for images of the event. They will also answer the questions using E-WKI which will be given the same data. They will perform each task with a different set of data which will be selected at random.

In each case the time taken by the citizen to make their decision will be recorded as well as their ability to identify the information correctly. Comparisons will be made using standard statistical methods in order to measure the differences between the two interfaces and whether E-WKI enables the Citizens to make a *better* impression of the event *faster* than with currently available data.

Following the evaluation, questionnaires will be used to assess the impression that the citizen made of the information and, again, interviews will be used to explore what informational views the citizen would have liked to have seen or what information the citizen feels would have assisted them in forming an impression of the event.

3.4.7. ER Professional Objective Evaluation

The objective evaluation for the ER Professional groups taking part in the study follows a similar line to that of the Citizen. Here, however, the ER Professional will be asked to make a broader assessment of the emergencies that are occurring.

The format of the evaluation is the same – the ER professional will be given access to E-WKI, the simple paper-based interface or just the comment data for the purposes of investigating three emergencies which are occurring simultaneously in a city. In both cases they will be asked:

- Where the events are occurring.
- What the severity of each event is.
- What type of event it is.

As with the citizens the ER professionals will be scored on the time it takes them to reach a decision as well as how well they were able to identify the relevant information. Again, statistical measures will be used to identify the differences between the two interfaces and whether E-WKI enables ER professionals to have a *better* understanding of what events are occurring *faster* than they do with the raw data that is processed by E-WKI.

Following the objective evaluation, questionnaires and interviews will be used to explore the informational needs of the ER Professionals in more detail. The same questionnaires will be used for both evaluation phases but the interviews in this case will focus on the informational requirements and how Collective Intelligence can assist each professional with collecting and understanding the required information.

3.4.8. Evaluation Report

The evaluation report will contain the full results of each of the user trials and the data collection phases of the experiment. In terms of the subjective evaluations the report will identify problematic areas of E-WKI both in terms of the system failures and interface failures. For example the subjective evaluation may identify that the users do not understand the process of tagging information and may additionally identify that incorrect tags are detrimental to the user.

The objective evaluation will be used to measure the success of the CI processing within WeKnowIt by performing a direct comparison between the CI view of the information and the raw view of the data that makes up the CI view. It is intended to determine how successful the added value provided by the CI processing is at assisting users when formulating an understanding of a typical emergency. In addition to this the report will detail any failings of the system from the perspective of the users of the system and how WeKnowIt can focus its efforts on supporting the kind of information that both citizens and ER professionals require in order to effectively make decisions and understand events.

4. Consumer Social Group Scenario

The focus in this particular case study is on travel or one-day cultural trip events. The scenario consists of different sub-scenarios:

1. **Travel preparation:** During the travel research, a user identifies a location of interest (POI, or Place). The prototype will at that point provide relevant information about the location. The user has the option to bookmark the location, or POIs within the selected location as a favourite. In this scenario we will leverage in particular the collective, social and media intelligence to offer the user a travel exploration experience that allows the user to collect essential information when preparing a trip or determining a holiday destination. The user activity is monitored and used in the second sub-scenario to personalize the mobile experience.
2. **Mobile guidance:** During a cultural trip or a travel a user uses his mobile, can take pictures and record videos with their mobile phones, which get added to the community area. Users can add notes or just rankings to those images and videos they upload. This will help other members to decide where to go, where to eat, etc. Also, the user may look for new events or point of interest using his mobile. The system provides him recommendations based on the user profile analysis in connection with his actual location, and suggests to the user events.
3. **Post travel experience** After or during the travel a user is given support in annotating their travel photos. During the Post travel experience, the personal intelligence, the mass intelligence and the media intelligence provided by WeKnowIt's layers to suggest annotation terms for photos and recall names of Places visited during the trip.

4.1. Scenario Definition

In the next sections we present the steps in which each sub-scenario is split and we describe the events that occur during the steps in each sub-scenarios.

4.1.1. Sub-Scenario Travel preparation

The prototype provides the following steps during the travel preparation phase:

- A. A user logs on to the E-WKI web-site to explore various locations
- B. The user carry out a search for a particular location, for instance "Barcelona"

- C. There are several locations, events, and/or points of interest that go by that name. Therefore the user is asked to disambiguate which location is meant by the user.
- D. For the location of interest, a location page is opened. The user can then explore various points of interest related to that location, upcoming events, or dig up more background information. During any of these actions they can also bookmark the item as a favorite.
- E. The user logs out, or performs another search.

4.1.2. Sub-Scenario Mobile guidance

In this section, we present the steps of the mini-scenario which are being developed for the first prototype. These steps are mainly related with the user's searches and the WKI's recommendations of POIs and Places. However, for the first prototype, WKI will not make recommendations based on user's GPS position or profile, due to the fact that characterization and personalization algorithms are still being analyzed. The late incorporation of Yahoo and TID to the WKI project, the mainly partners involved in the CGS development, has caused that the recommendation of Events has had to be postponed for the second prototype due to the lack time for the development. The first prototype provides the following steps during this phase:

- **MG1:** During the visit, a group of users, Peter, John and Mary are going to see a spectacle in Barcelona. Suddenly, it is starting to rain and Peter looks for alternative POI to go to, by accessing WKI with his mobile.
 - WKI suggests visiting the "La Sagrada Familia"
 - John reads the opinions, sees the pictures and the information about the "La Sagrada Familia" on WKI
- **MG2:** Then the group decides to have dinner. John searches restaurants in the Place selected.
- **MG3:** Mary does not like the restaurant which Joseph has chosen therefore she looks for other restaurants. WKI suggests some restaurants. Mary reads the information about this. Based on this information, she makes her choice.

4.1.3. Sub-Scenario Post travel experience

After a trip the user has collected various information, such as, photos, list of points of interest visited, and GPS coordinates (either in the form of sparse points or more detailed trail, depending on device). The user is given support in annotating their photos, both the ones uploaded during the trip and the ones uploaded after the trip.

4.2. Evaluation

The evaluation for the first prototype of the WKI system will therefore consist of various evaluations: One for the users using the desktop device during the stages of **Travel preparation**, a second using the mobile device during the **Mobile guidance stage**, and finally when they return from the trip and upload their photos and experiences. YAHOO will support evaluation activities during the stages of Travel preparation and Post travel experience and TID will carry out the evaluations activities during the stage of Mobile guidance.

The procedures will consist of a number of tasks for the participant to carry out with the system, how performance of the system will be measured with regard to the task. The follow up questionnaires and suggested interview questions are contained in the appendices.

4.2.1. Sub-Scenario Travel preparation

4.2.1.1. Evaluation Tasks

We will prepare a number of search-and-explore tasks for the participants of the experiment, and use the logging facilities of the prototype to monitor the users' behavior. Prior to, and after the experiment, we will ask participants to complete a questionnaire.

4.2.1.2. Success Measures

- Effectiveness: What is the success rate of the participants when trying to locate particular information, or to perform a certain task.
- Efficiency: For the successfully executed task, how long did it take them to perform that task, and what are the reasons for failure.
- User Satisfactory: To what extent are the users satisfied with the system.

4.2.2. Sub-Scenario Mobile Guidance Stage

4.2.2.1. Evaluation Tasks

The main tasks that the user should evaluate are presented below.

1. Search and Review recommended Places (for example Barcelona)

The user makes a Place query. WKI returns a list of recommended Places that match the query.

2. Retrieve information about a Place

User selects one Place on the recommended Places list. WKI shows him the Place information and also shows him clustered information for several POIs in the Place as well.

3. **Search and Review recommended POIs** (for example restaurants, see sights and monuments)

The user makes a POI query. WKI returns a list of recommended POIs that match the query.

4. **Retrieve information about a POI**

User selects one POI on the recommended POIs list. WKI shows him the POI information.

4.2.2.2. **Success Measures**

In this section, we have listed the measures of success expected for the tasks presented above. The results expected have been defined using pre-test cognitive walkthrough. In mobile device, the output information should be rendered according to limitation of size in mobile devices.

Search recommended Places or POIs

The cognitive walkthrough should confirm that it is possible to make queries about Places, POIs and get the list of Places or POIs recommend by E-WKI.

- **Success:** User is able to get the list of recommended Places, or POIs and this constitutes a good list. That means, the list should contain all Places, or POIs that match the query and the information showed has been rendered according to limitation of size.
- **Partial Success:** User is able to get the list of recommended Places, or POIs and this constitutes a fair list. That means, the list should not contain all Places, or POIs that match the query or the information showed has been rendered according to limitation of size.
- **Partial Failure:** User is able to get the list of recommended Places, or POIs and this constitutes a poor list. That means, the list contains Places or POIs that are not relevant to the query.
- **Complete Failure:** User is unable to get the list of recommended Places, or POIs.

Review recommended Places or POIs

The cognitive walkthrough should confirm that it is possible to navigate though the list of recommended Places, or POIs to its overview and their relevant information is displayed.

- **Success:** The information displayed to the user constitutes a good overview and has been rendered according to limitation of size.
- **Partial Success:** The information displayed to the user constitutes a fair overview to the user.

- **Complete Failure:** The information displayed to the user constitutes a poor overview.

To guarantee the success of these tasks, user should be interviewed and should fill a questionnaire. The purpose of this questionnaire is to identify the difficulties that users have found during these tasks and if the functionality developed is useful for their needs.

The other objective of this questionnaire is to determinate why the user has been unable to complete any of the tasks and the possible problems related with it.

4.2.3. Sub-Scenario Post travel experience

4.2.3.1. Evaluation Tasks

- Recommend annotations for photos

4.2.3.2. Success Measures

The performance (and hence evaluation) depends on the extensiveness of data collected through the pre-travel and mobile-travel scenario and cannot be evaluated until the second round of prototypes.

- User satisfaction
- Precision / recall performance of the annotation suggestion.

4.3. Scenario Constraints and Requirements

4.3.1. Technological Constraints

E-WKI is accessed over both mobile devices and from desktop computers. A standard Internet connection and class A browser is required. For the pre- and post experience, Yahoo! will carry out a QA test, as is standard procedure for any publicly accessible portal.

The range and diversity of marks and models of mobile devices is much wider than PCs. To guarantee E-WKI works correctly, the mobile guidance functionality has been developed until the time of this writing which supports mobile devices with the minimum feature of a web browser and GPS-enabling capability. Therefore it will be sufficient to perform a cognitive walkthrough on the main browser types to ensure that the system functions correctly across these browsers.

4.3.2. Temporal Constraints

Although, the scenario defined above occurs over different places and times, to consider them is not strictly necessary for the first prototype evaluation, because the temporal changes and the changes of places are not part of the interaction with the system.

4.4. Target user groups

In the context of Consumer Group Case scenario, we can't identify different roles for the users. The same users that use their PCs to collect the information when they are preparing a trip might use their mobile to take pictures and to search during a travel. And their roles aren't different either. The only difference between the groups of the users is the type of device (mobile or desktop PC) they are using to test the scenario and their location during the test. Therefore, we consider two groups of users: One group is represented by the users using the **desktop device** and the other one by the users using a **mobile device**. For the development of the first prototype we have considered that WKI exclusively addresses users using the high-end models of mobiles, due to the limitations in the development time.

The group of users using the **desktop device** can be simulated in the lab studies. However, the group of users using the **mobile device** should be essentially users located outside the labs and with access to mobile devices capable of calculating the user's GPS position and to access the Internet. Nonetheless, the user's GPS position could be simulated for users in the labs studies for the evaluation first prototype.

4.5. Evaluation Work Plan

The work plan for whole process of evaluation is summarized in the following table:

	M16	M17	M18	M19	M20	M21	
Desktop interface. Partner Responsible Yahoo							
evaluation work plan	done						
task definition				targeted			
evaluation platform					targeted		
evaluation experiment						targeted	
analysis of results							targeted
Mobile Interface Partner Responsible TID							
evaluation work plan				targeted			
task definition				targeted			
evaluation platform					targeted		
evaluation experiment						targeted	
analysis of results							targeted

Figure 2: Evaluation Work Plan for the CSG evaluation

The evaluation of E-WKI includes recruitment of participants, for the evaluation of CSG case we have planned two possible lines of action:

1. Evaluation volunteers who would be willing to participate could be recruited amongst partner organizations involved in the development of this case. YAHOO's employees will be recruited to support the evaluation activities during the stage of travel preparation and TID's employees will specially be encouraged to carry out the evaluation of the mobile guidance phase tasks, using high-end models of mobiles. In the case of TID's employees will not be able to carry out the evaluation on their personal mobile devices, TID will provide this type of mobiles to their employees. The minimum required number of users to be recruited is dependent on the number of user required to collect sufficient data for the personalization phase.
2. On the other hand, if the CSG evaluation should be preferentially realized by the same group of users, evaluation volunteers might be recruited amongst the members of the consortium, and it might be performed during the next WeKnowIt's meeting. The participants in the next WeKnowIt might evaluate the tasks related with the Travel preparation before the meeting using their PC's for a week. Then, the mobile guidance phase tasks might be evaluated during one of the meeting session by the members of the consortium using this owns mobile devices.

For both possibilities of the first prototype evaluation, E-WKI will just let users make plans to travel (or during the travel) to "Barcelona" or "Madrid", because of this is the type of information available in WKI at the time of writing this.

After the evaluation, the results will be analyzed and reported to the member of the consortium internally. The results of the evaluation procedure will be presented in an evaluation report. Concerning to the writing of evaluation report, TID will be the responsible partner of the writing of mobile guidance evaluation report, and Yahoo will support the writing of report for travel preparation tasks.

The following picture shows what will be the architecture used during the evaluation period.

5. Conclusions

The evaluation procedure for the WKI system has been outlined in the previous sections. For the **Emergency Response Case**, two different evaluations will be used: One for Sheffield citizens, and one for ER Staff. In both cases the evaluation process a task based to success measures defined individually for each assessed functionality/task. To determine whether the WKI system has met the required success measures a video recording will be made for each user which will be analyzed after the evaluation session. Questionnaires and informal interviews will also be used to determine the user reaction to the system and the interface.

In the context of **Consumer Group Case**, we have considered two types of user groups for the evaluations: One group is represented by the users using a desktop device and the other one by the users using a mobile device. However, the only difference between the groups of the users is the type of device (mobile or desktop PC) which they are using to test the scenario and their location during the test.

The correct functionality of the **Consumer Group Case** first prototype will be evaluated by performing functional tests, and getting the user's feedback with specific evaluation questionnaires as well. The purpose of these questionnaires is to identify the difficulties that users have found during these tasks and if the functionality developed is useful for their needs. In the functional tests, the test subjects will execute a sequence of tasks, and will assess the degree of success. For the performing of CSG functional tests, either the test subjects could be recruited amongst the members of the consortium and the evaluation could be performed during the next WeKnowIt's meeting, or they could be recruited amongst the employees of the main partners involved (Yahoo and TID). In the case, the minimum required number of users to be recruited is dependent on the number of user required to collect sufficient data for the personalization phase.

6. References

- [1] WeKnowIt Annex I – “Description of Work”, ver.1, FP7-215453, 26 Oct. 2007
- [2] A. Dix, J. Finlay, G.D. Abowd and R. Beale, “Human-Computer Interaction”, Prentice Hall, 1993.
- [3] 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.
- [4] Likert, Rensis "A Technique for the Measurement of Attitudes". Archives of Psychology 140, pp. 1–55, (1932).
- [5] J. Nielsen and R. Molich, “Heuristic evaluation of user interfaces”, Proc. ACM CHI, Seattle, 1990, pp 249-256.
- [6] WeKnowIt “D7.1 Consumer and Emergency Response Use Case Initial Requirements”, Sep. 2008
- [7] WeKnowIt “D7.2 Emergency response and consumers’ social group case study design and specification”, Feb. 2009
- [8] WCAG 2.0 Guidelines, <http://www.w3.org/TR/WCAG20/>

A. Overall questionnaire for both scenarios

These questions should be answered with a five-point Likert scale.

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

B. Emergency Response Case

B.1. Subjective Task Sheet for Sheffield Citizens

In this experiment you are a citizen reporting a flooding incident to the Emergency Response system. You will upload a photo you have taken to the system.

The facilitator will assist you if you need help but you should try and complete the tasks as best you can *without assistance*. Do not worry, however, if you are unable to complete some of the tasks - the system is just a prototype and may not be fully functional.

If you do not have a suitable mobile device then you will be provided with one and given a short demonstration of how the device functions. If you have a suitable device the experimenter will explain how the experiment will work. The facilitator will tell you if the device you have will be suitable for this facilitator.

If you do not understand some of the tasks then please ask the facilitator for clarification or, if at any stage during the experiment you are confused or do not know what to do, do not hesitate to ask the facilitator for assistance. The test and the subsequent discussion should take from 10-20 minutes.

If you have any questions about the experiment then please ask them now.

Mobile Tasks

You should attempt to carry out all these tasks on the mobile device.

1. Take a picture of the event

Take a picture of the event using your mobile device.

2. Log In

Using the provided credentials, please log in to the system.

3. Upload an image

Once you have logged in you can upload the photo you took to the system and follow the screens to complete the upload process.

Wrap Up

Your facilitator will now ask you a few questions about what you have just done and you will be asked to fill out a questionnaire. There will then be a

few more questions about the system in general and then the experiment will be complete.

B.2. Questionnaire for Sheffield Citizens (Subjective Evaluation)

For the mobile interface

5-Scale Likert Questions:

1. I found it straightforward to log in to the system.
2. It was clear from the interface how I should upload an image.
3. It was simple to tag an image.
4. I could understand what I was able to do with the mobile interface.
5. I think I could use this interface in a time-critical situation

Free Response Questions:

1. What did you like about the mobile interface?
2. What did you dislike about the mobile interface?

B.3. Task Sheet for ER Staff (Subjective Evaluation)

Prior to starting the experiment the facilitator will give you brief instructions on how to use the system and provide you with your login credentials. In brief the experiment will involve you reviewing information captured by citizens, allowing some of this information to be seen by the public, reviewing the information received and reviewing new content. All of this occurs within the context of a flooding event.

You should attempt to do each of these tasks without assistance but please ask the facilitator if you get stuck doing anything or any part of the task is unclear. This is a prototypical system so do not be surprised if some parts do not work or produce strange results. The test and the subsequent discussion should take from 10-20 minutes.

Tasks

Please carry out the following tasks with the system:

1. Log In

Log into the system, using the credentials you have received.

2. Review all the information

The system will show you all the information relating to events in the City. Please take a moment to examine this information.

3. Look specifically at one of the events

Choose one of the events and take a moment to examine the information that specifically relates to this event.

4. Approve some images for public viewing

Choose some of the images from the flooding event and change them so that they can be viewed by members of the general public.

Wrap Up

The facilitator will now ask you a few questions about what you have done. Following this you will be given a questionnaire to fill out and there will be a few more questions and the evaluation will be complete.

B.4. Subjective Questionnaire for ER Staff

Since there will be fewer ER Staff taking part in the evaluations, and the critical parts of the interface are more information based rather than task based, less weight should be placed on the questionnaire. The post-experimental interviews should explore the suitability of the system overviews to support the tasks that the ER staff typically have to carry out.

5-Scale Likert Questions

1. The initial system overview clearly presented the key information for the city-wide events.
2. The specific flooding overview presented the key information clearly.
3. It was straightforward to approve images for public viewing.
4. The overview of the event information received was useful.
5. It was simple to distinguish new content from old content.

Free Response Questions

1. Identify three things you liked about the interface
2. Identify three things you did not like about the interface

Suggested Interview Questions

1. What did you like about the way in which the overview information was presented?
2. What, if anything, was missing from the overall overview screens?
3. What, if anything, was missing from the specific overview screens?
4. How do you think, seeing the information presented by E-WKI, you would have acted differently from your standard procedure?
5. In what ways do you think the E-WKI has improved your understanding of the flooding event?

6. What tasks would you like to be able to carry out with E-WKI, in addition to ones you have seen in the prototype?
7. Which parts of the interface were unclear to you?

B.5. Objective Data Collection Tasks

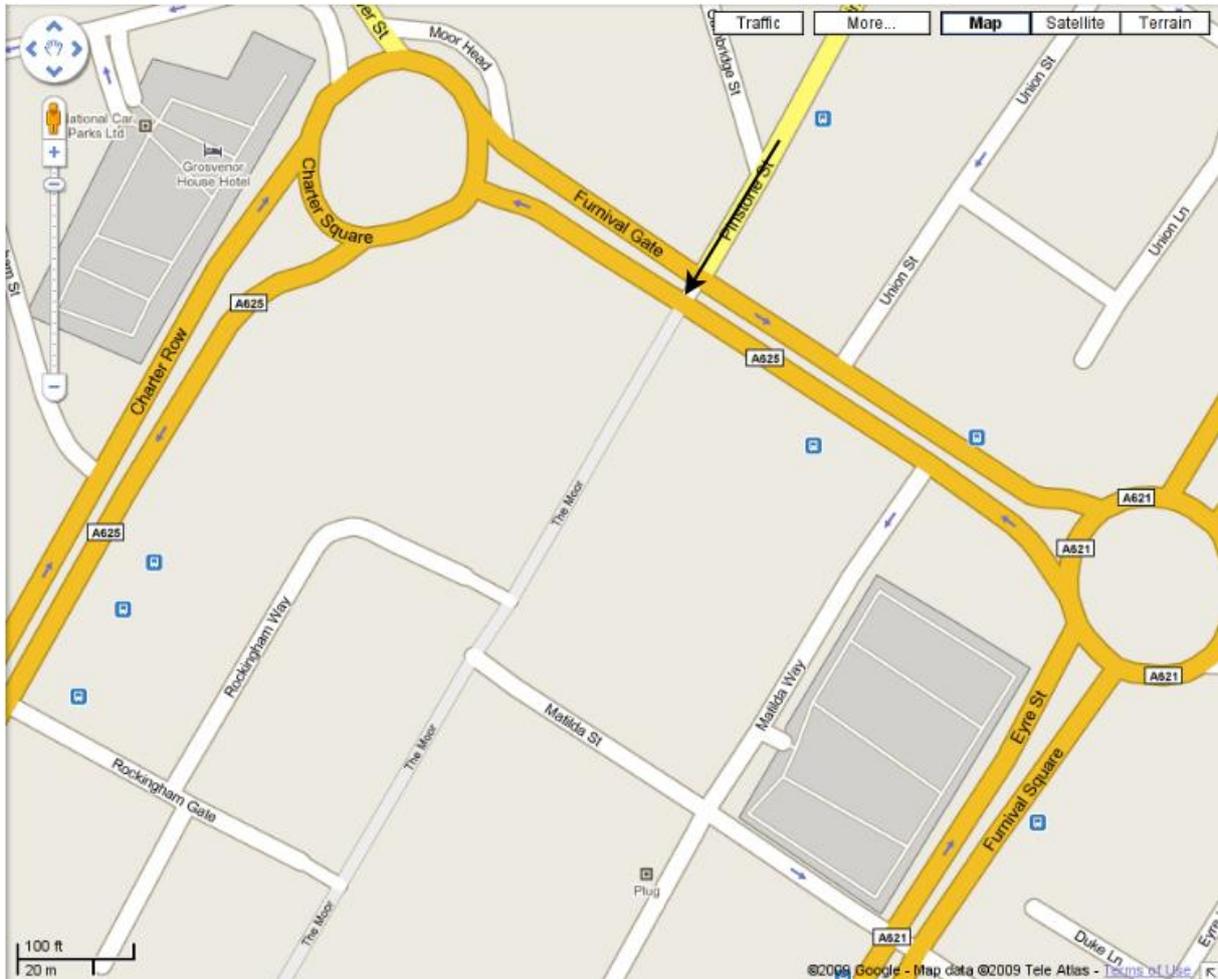
The overall task for the data collection phase will be given to the participants as follows:

You are about to take part in a simulation of an emergency somewhere in Sheffield City Centre. Your task is to follow the given directions and when you reach the emergency location upload 5 different images of the “event” and tag the images using the interface provided. In addition you should send a short text message describing the event to the number provided. Please make sure that you arrive at the allotted place at the given time as this information is important towards supporting the upload process. Please also ensure that you understand how to use the device supplied to you – the facilitator will give you a short demonstration on how to use it correctly.

In addition to this general information each event participant will be given the specific details of the event and the location information as follows. Multiple arrows are used to indicate the different directions that the participant should approach the event from – each participant will be assigned to a single direction.

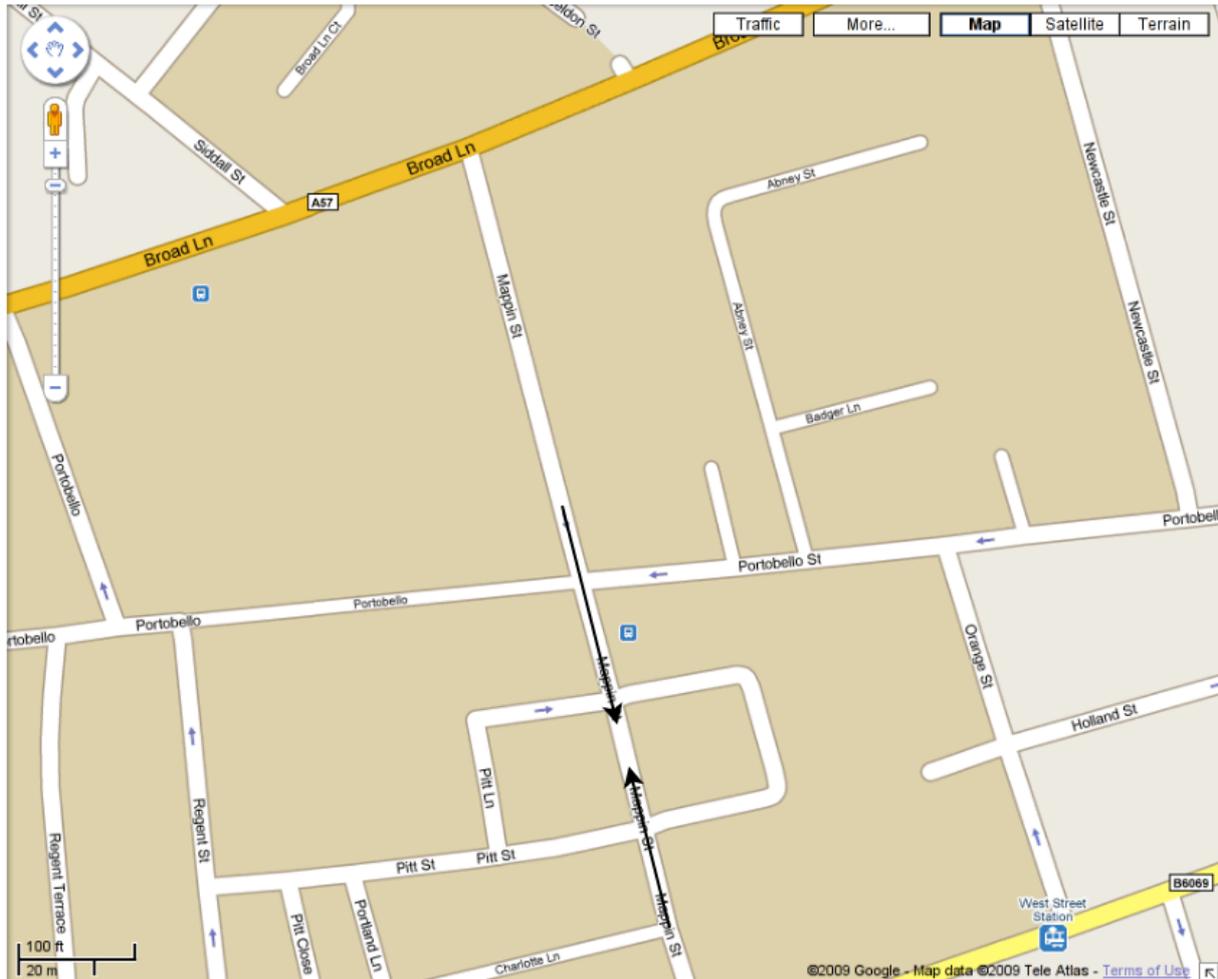
Fire – Low Severity

You witness a fire in a bin at the given location. It seems in control and will probably go out in the next few minutes.



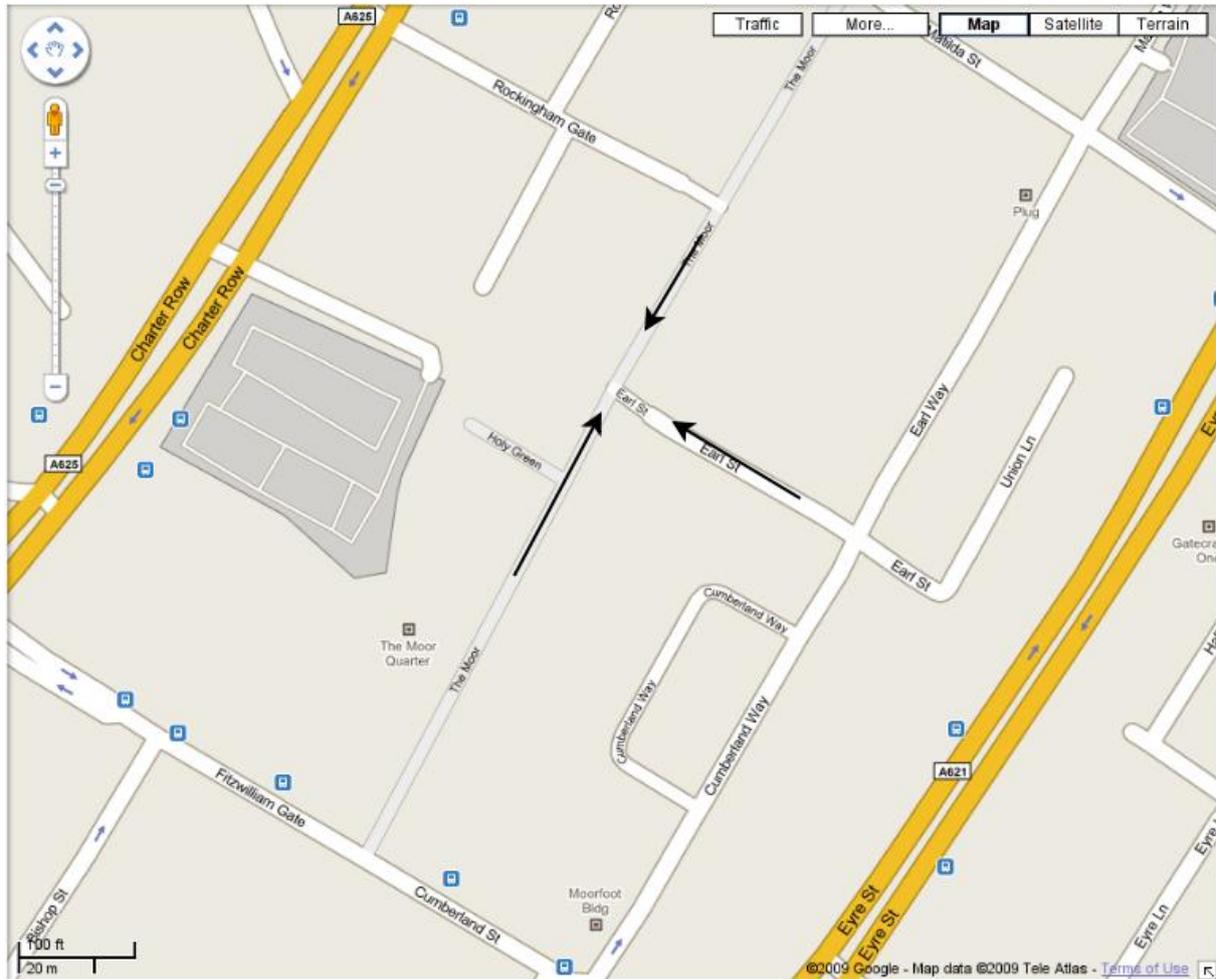
Fire – Medium Severity

You witness a fire in the kitchen of a block of flats. The fire appears to have been caused by a toaster but there is a danger that it could spread.



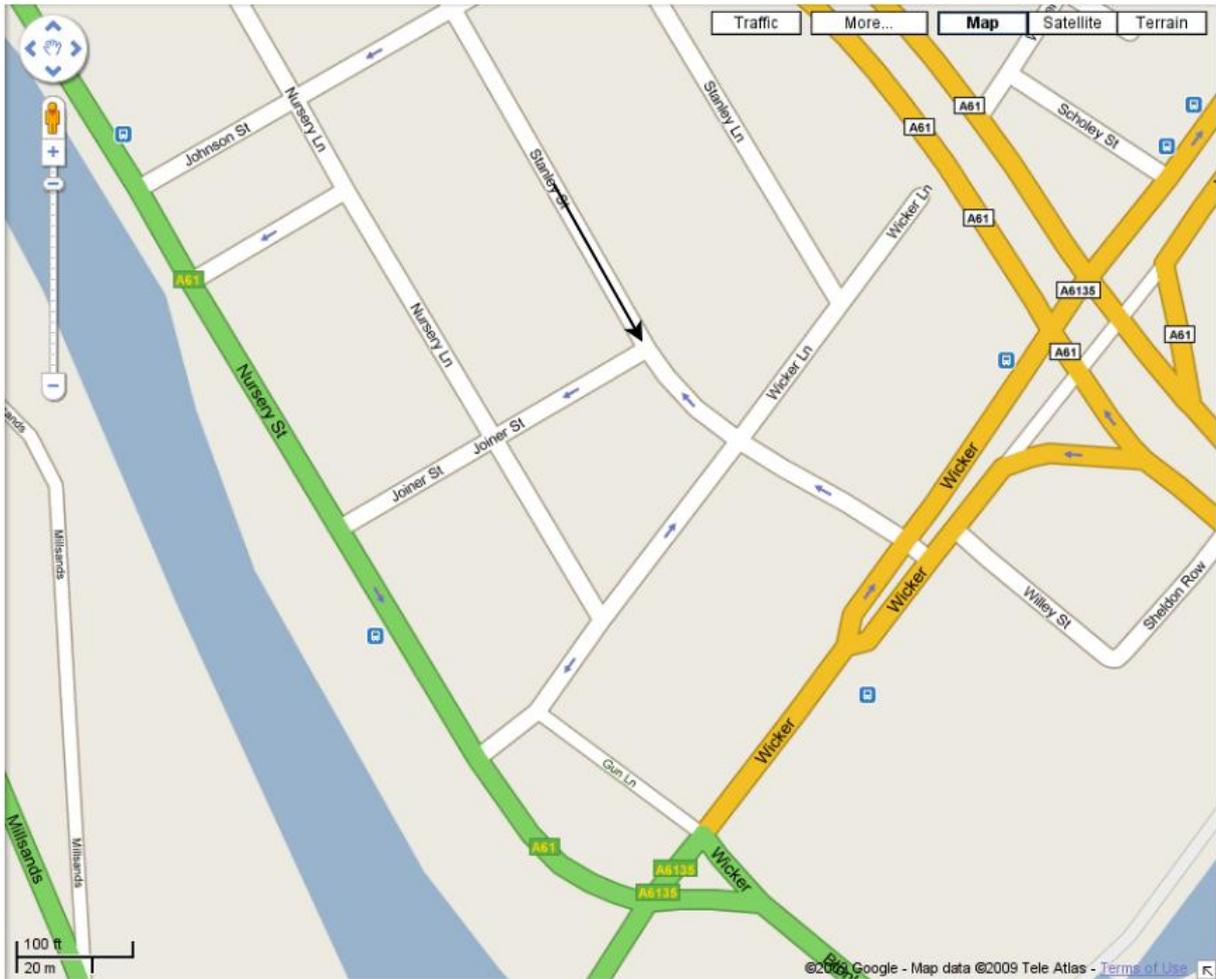
Fire – High Severity

You witness a large fire in the Boots store in the Moor. The fire is substantial and looks like it's out of control.



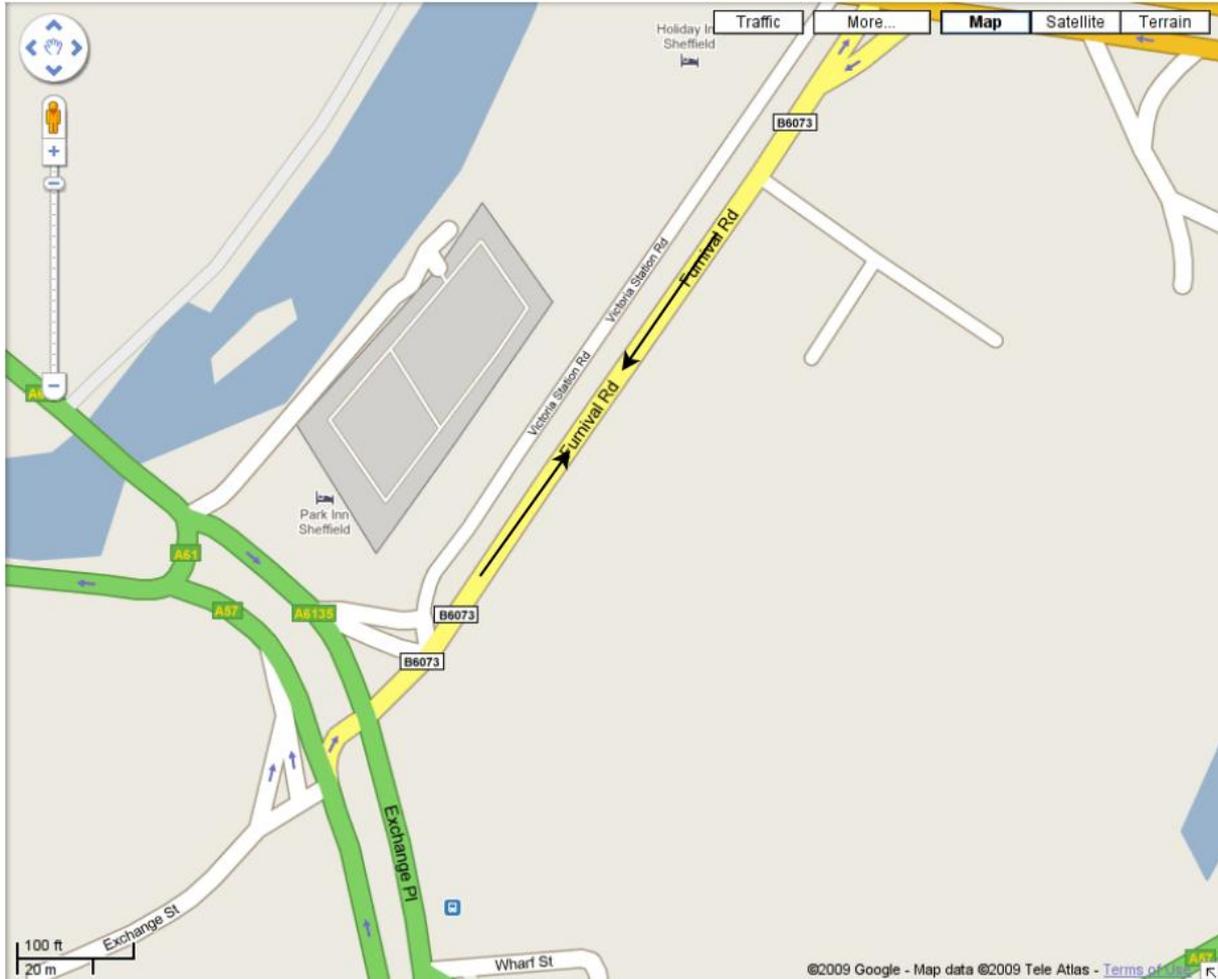
Flooding – Low Severity

You see that the T-Junction at Stanley Street is flooded. It is a minor road and the flooding does not look severe.



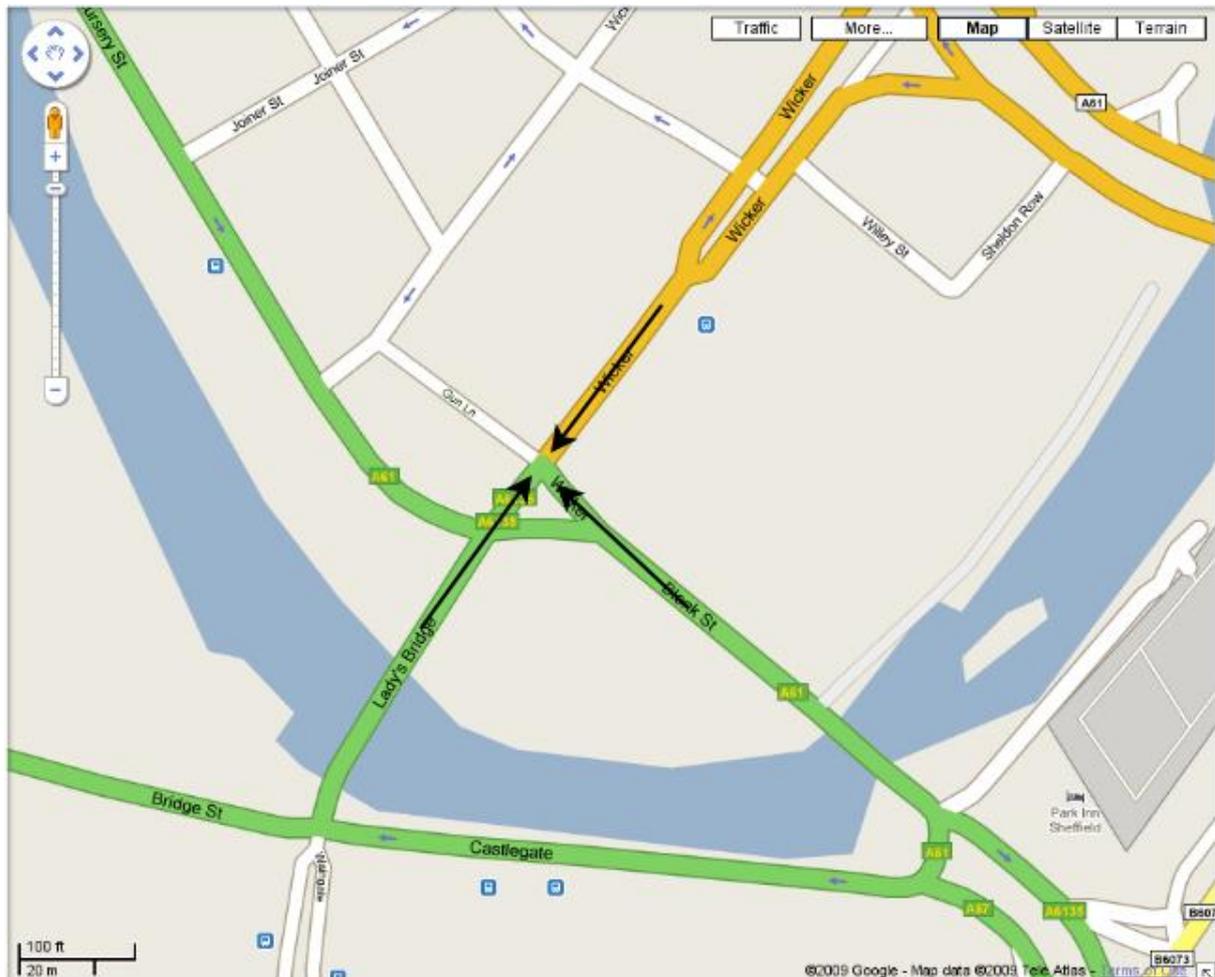
Flooding – Medium Severity

You see that Furnival Road is severely flooded – whilst this isn't a major road the flooding is serious and could get worse over the course of the day.



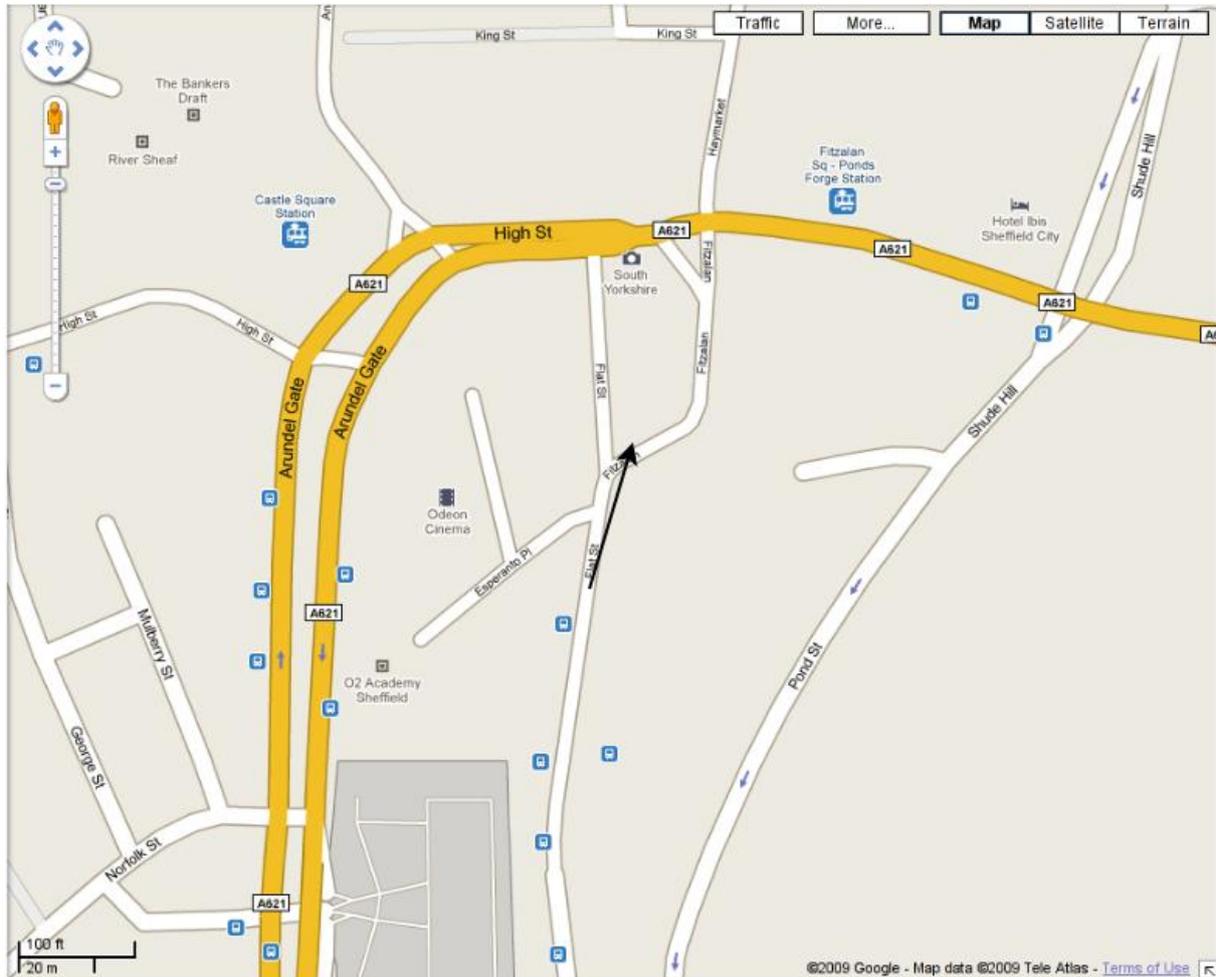
Flooding – High Severity

You see that the Wicker is severely flooded. There is a lot of water and a strong indication that the flooding will get worse. There are cars and people trapped in the flood waters.



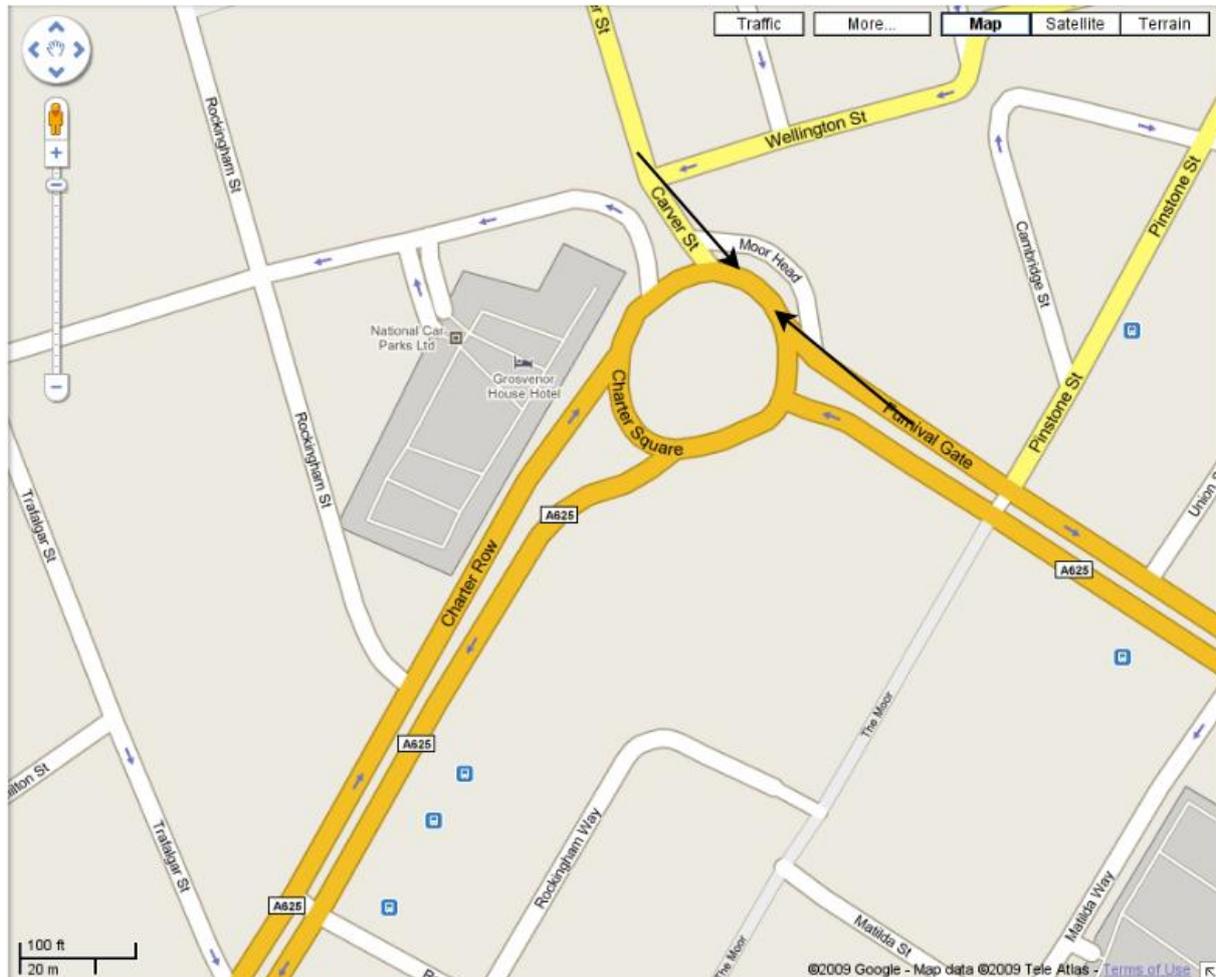
Tree Fell – Low Severity

You see that a tree has fallen over in the square at Flat Street. It is not blocking a road although there are some minor injuries



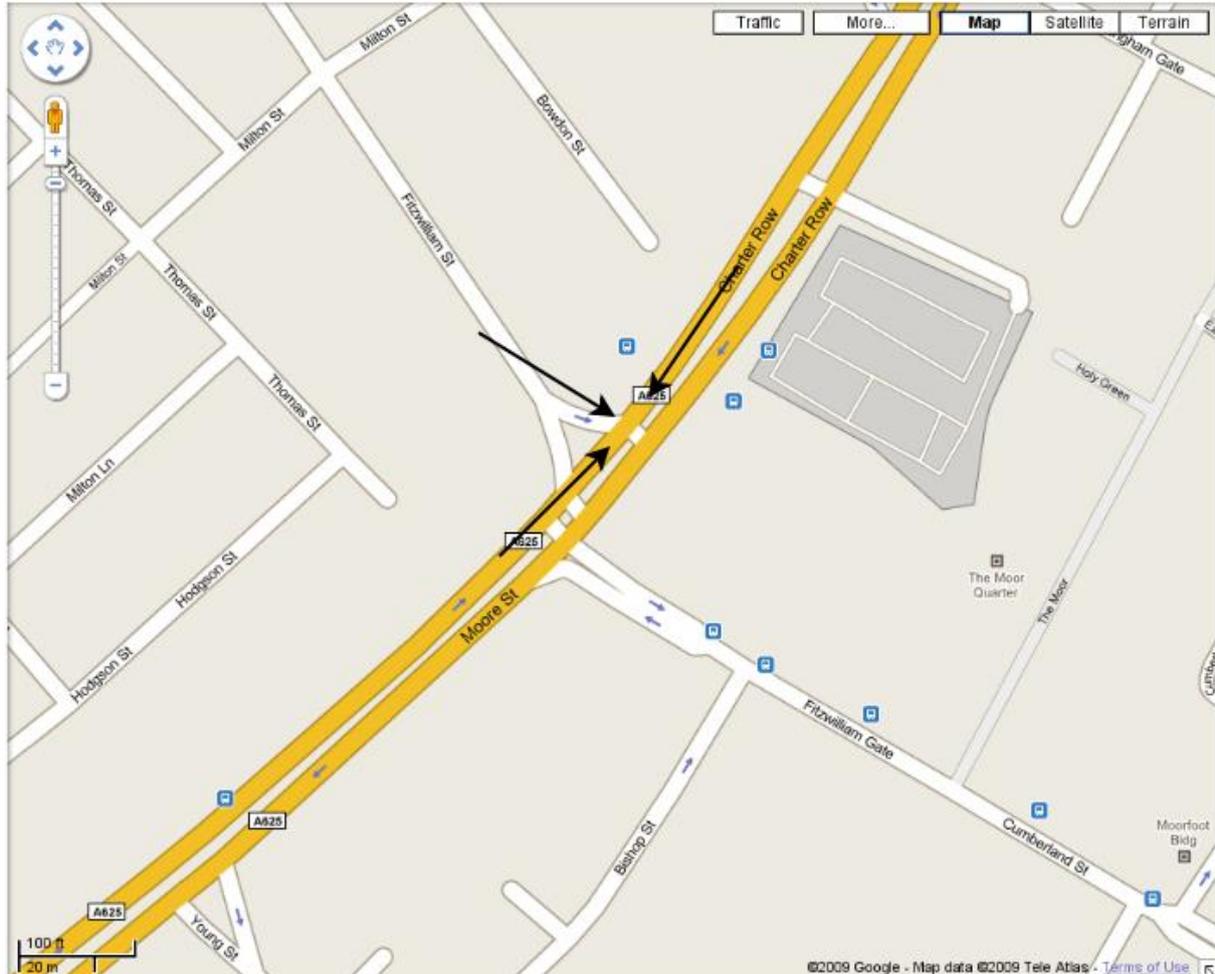
Tree Fell – Medium Severity

You see that a tree has fallen over on the Charter Square roundabout. There appear to be no injuries but the road is blocked and traffic is building up.



Tree Fell – High Severity

You see that a tree has fallen over and is completely blocking Charter Row. In addition the tree has fallen on two cars, one of which appears to be carrying passengers, the other of which appears to have no driver.



B.6. Task Sheet for Citizen Objective Experiment

In the following experiment you will be given three interfaces which you can learn about an event which occurred during the day in the city. You witnessed the event from a long distance and want to use the interfaces to learn more about the event. Specifically you would like to know

- What the emergency was
- Where it took place
- How severe it was

In one interface you are given some pieces of paper each containing information corresponding to three emergencies occurring in the city during the day. In another you will be given just textual information. The third interface will present the same information in an interactive online interface. In each case you must process this information to collect the

details relevant to the event you saw and answer the questions above. You will be given a short training session using a different data set in order to learn how to use the interfaces.

You can take as much time as you wish to complete the experiment but you should keep in mind that you do not have to view every single piece of information to make your decision but should view enough information in order to feel confident that you have made the correct decision regarding the three questions asked of you.

After the experiment there will be a short questionnaire to fill out and the facilitator will ask you some questions about how you used the interfaces to complete the tasks.

B.7. Task Sheet of ER Professional Objective Experiment

In the following experiment you will be given two interfaces which you can use to learn about three emergencies which occurring simultaneously throughout the city. Specifically you would like to know

- What the emergencies were
- Where each took place
- How severe each was

In one interface you are given pieces of paper containing information relating to the three emergencies occurring in the city during the day. In another interface you will just be given text comments relating to the events. The third interface will present the same information but used an online interface to display the information. In each case you must process this information to collect knowledge relating to the events and answer the three questions above. You will be given a short training session using a different data set in order to learn how to use the interfaces.

You can take as much time as you wish to complete the experiment but you should keep in mind that you do not have to view every single piece of information to make your decision but should view enough information in order to feel confident that you have made the correct decision regarding the three questions asked of you.

After the experiment there will be a short questionnaire to fill out and the facilitator will ask you some questions about how you used the interfaces to complete the tasks.

C. Consumer Social Group Questionnaires

GROUP: _____
 NUMBER: _____

USER

Please mark your answer to each question with an X to the correspondent square, where the values are:

Strongly disagree

Strongly agree

1	2	3	4	5
---	---	---	---	---

Specific questions

	1	2	3	4	5
I find useful the information showed about Places, POIs and it helped me to prepare (or during) my trip.					
I find useful that the system showed me the clustered information for several Places with the same name I was searching for					
I find useful that the system showed me together all the Place information (photos, general information, etc) aggregated from different sources					
I find useful that the system showed me the clustered information for several POIs related with a Place.					
I find useful that the system showed me together all the POI information (photos, general information, etc) aggregated from different sources					

Feedback questions

What problems did you find while using the system?

Suggestions to improve the system