Mobile Evaluation: What the Data and the Metadata Told Us

Sonja Pedell[1], Connor Graham[2], Jesper Kjeldskov[3] and Jessica Davies (nee Smith)[4]

[1] Department of Information Systems, Melbourne University, pedell@acm.org
[2] Department of Information Systems, Melbourne University, cgraham@unimelb.edu.au
[3] Department of Computer Science, Aalborg University, jesper@cs.auc.dk
[4] Department of Geomatics, Melbourne University, jcsmith@sunrise.sli.unimelb.edu.au

Abstract

Evaluating mobile applications to identify usability problems presents a unique set of challenges. Not only is it difficult to capture data on an application that is inherently mobile, but generating an authentic environment of use is also problematic. This paper compares two "traditional" user-based approaches to evaluate a mobile system: one laboratory-based and the other in the field. These data serve as a basis for the primary focus of this study: the effectiveness of 'metadata' ,generated from 'rapid reflections'. These data were collected by the evaluators after each day of evaluation in order to investigate the quality of metadata in relation to the data of the two evaluation methods. The study also found that the laboratory study identified typical usability problems with the system at a more detailed level whilst the field study identified characteristic problems of mobile use. The metadata findings summarised the major findings in a useful way, but generally were less specific and reflected subjective theories of individual researchers.

1. Introduction

There exist many approaches to evaluation in Human-Computer Interaction. These include user-based evaluations such as usability studies in specialist laboratories, expert-based evaluations such as heuristic review, and theory-based evaluations such as the Keystroke Model. A similar proliferation of approaches exists in Mobile HCI (Kjeldskov & Graham, 2003). In both fields, techniques of data collection and analysis vary considerably across approaches. For example, an expert-based evaluation may involve collecting data from usability experts and analysing it to identify common themes whereas a user-based evaluation may involve using quantitative data to determine the efficiency of task completion using statistical approaches.

When designing evaluations key decisions must be made regarding method, technique and analysis. Although there is a strong body of research in human-computer interaction (e.g. Gray & Salzmann, 1998; Henderson et al., 1995; Karat et al., 1992) regarding the appropriate choice of method, data collection and analysis technique, much less research has been conducted examining and comparing methods and techniques for mobile system evaluation (Kjeldskov & Skov, 2003). The move beyond stationary settings has created new challenges for the evaluation of useful and usable systems (see e.g. Luff & Heath, 1998) and has reinforced the discourse on laboratory versus field testing. This discourse promises to become more interesting with a recent review of mobile research methods showing that the majority of research conducted to evaluate mobile systems uses laboratory experiments over field studies (Kjeldskov & Graham, 2003).

Decisions taken regarding the choice of evaluation method(s) and technique(s) of data collection and analysis within the field of human-computer interaction are often pragmatic both within and without a research context. For example, Nayak et al. (1995) used the following five criteria to enable workshop participants to rate evaluation techniques: positive effect on team acceptance, amount of time required to use the technique, degree of special expertise required, ease of translation into design changes and probability of introducing bias. Notably, three of these criteria are pragmatic and "classic" principles of data quality, such as reliability and validity are only indirectly referred to in the last criterion. Other researchers highlight the importance of such principles. For example, Gray & Salzmann (1998) describe validity as the core issue in judging usability evaluation methods, stressing the importance of conclusion validity through a strategy of triangulation. However, in a time-critical setting, principles of data quality may be emphasised less and analysis may be reduced to a discussion of the results. The focus of this paper is not only to reflect on the time effectiveness and "quality" of methods used, but also the time needed for analysis of different data.

With the rise of more qualitative approaches in human-computer interaction (e.g. Millen, 2000; Braiterman & Larvie, 2002), questions of pragmatism, validity and reliability become even more important. Ethnographic approaches have been made more pragmatic through the development of rapid ethnography (Millen, 2000). However, questions of data validity and reliability remain given that the data is often "messy" and based on a small user sample and are the responsibility of a few highly involved researchers to interpret and present. Some even argue questions regarding reliability and validity are simply not relevant for naturalistic enquiry. For example, Guba & Lincoln (1989) describe principles of *credibility* and *transferability* over *internal validity* and *external validity*.

More qualitative approaches offer exciting possibilities for mobile human-computer interaction. The question still remains regarding how we establish the strengths and weaknesses of one method or technique over another? Gray & Salzmann (1998) advocate experimental approaches, whereas Olson and Moran (1998, p.295) criticize this, advocating "a consideration of a broader range of evaluation methods for UEM [usability evaluation method] than the "narrowly focused experiments" advocated by Gray & Salzman".

In this paper we firstly compare two data sets generated from an evaluation of a mobile route-planning system across two empirical (Gray & Salzmann, 1998) or user-based evaluations: a 'standard' set of video data and a more informal set of results or rapid reflections, gathered from interactions among evaluators of the same system. We use the term 'rapid reflection' in this context in a similar fashion to rapid ethnography (Millen, 2000). It is defined as focused and pragmatic discussion and consideration of collected data by researchers.

We describe the results collected from users as user data and the data collected through rapid reflection as metadata. This is labelled metadata because it describes the user data from the two empirical studies. We also compare the results generated from the different evaluation settings within each data set from the same evaluation. In this analysis we do not claim statistical power but the deployment of a rich, qualitative approach to explore the relationship between the data and the metadata. This permits us to draw some conclusions concerning the most cost effective data collection method and analysis method.

The objectives of the research are described in more detail in the next section.

2. Research aim and objectives

The primary aim of this paper is to understand how evaluation techniques can be combined to produce effective evaluations for mobile systems. More specifically, the primary objective is to understand the differences between two *data types* collected from a usability evaluation, in this case laboratory and field *metadata* and laboratory and field *user data*. Metadata are defined here as a series of observations on the user data collection process itself, including minutes of evaluator meetings and evaluator diaries. The meeting metadata were generated from rapid reflection on a day's work collecting user data. User data are defined as typical information collected during usability studies such as verbal protocol and videos of user behaviour.

The secondary objective is to understand the differences between the results collected from two *data settings*, in this case a field and laboratory setting. In generating these objectives, we were aware of the role of pragmatism in evaluating mobile systems. In particular, we wanted to understand how "expensive" each data type and data setting was.

Thus, we firstly aim to investigate the effectiveness of the kind of data that is often evolved during meetings in industrial usability testing settings for the evaluation of mobile systems. Secondly we aim to understand what laboratory and field studies can offer mobile device evaluation and to understand the kinds of phenomena each method is effective at evaluating. We regard this as important in informing the choice of evaluation and analysis approach for both academic researchers and industry specialists. The relationships between research objectives and data sets are shown in figure 1.

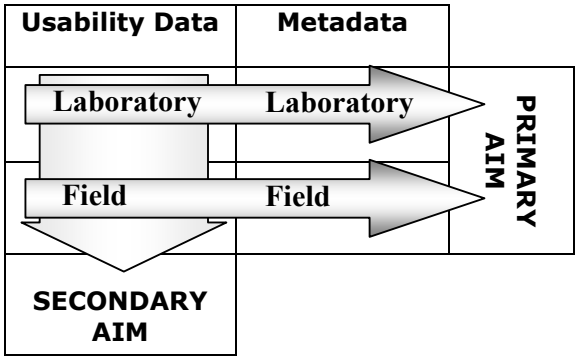

Figure1. Overview of research objectives and data sets

3. Background

These evaluations took place as part of the TramMate Project, running within the Department of Information Systems with the collaboration of Novell. The aim of this project was to develop a mobile service prototype to support the use of public transport systems in Melbourne, Australia. The initial design of TramMate was based on field studies on the use of transportation by business employees who, during a typical workday, have to attend appointments at different physical locations. From these field studies some key requirements were identified (Kjeldskov et al., 2003):

- Relating travel information to appointments;

- Providing route-planning information based on current location;
- Alerting the user about departure times;
- Providing access to information about walking distance and route changes.

In association with the Department of Geomatics at the University of Melbourne a location-aware trip planning prototypical service was evaluated parallel to the design of TramMate. This system mirrored many of the requirements for the envisaged design and is described in the next section.

4. The evaluated system

The application was designed for use on an iPAQ handheld computer equipped with a WAP browser. The device was connected to the Internet via a GPRS data connection. Using a handheld computing device (rather than a mobile phone) also facilitated the integration of a global positioning system (GPS) receiver for monitoring the position of the device. To ensure reliable positioning throughout the evaluation period, GPS positions were simulated within the system.

The application was designed to serve three functional processes with regard to public transport. These were accessible via the startup screen.

- Timetable Lookup: information about the tram timetable based on the input of stop numbers (origin and destination) and route numbers. This function was aimed at regular tram users who are very familiar with their route of travel. No maps are available within this section of the system.
- Plan Trip: information about the whole route (containing route descriptions and maps) based on the input of suburb and street corners of origin and desired destination. Users were also presented with an option to enter an arrival time or departure time for their journey. From each screen within this function, it was possible to view a visual representation of the relevant portion of the journey on a map.
- Determine Route: information about the whole route (containing route descriptions and maps) based on the input of the street corner of the destination and the suburb. The system determined the user's origin location via a (GPS). Maps were also available for components of the journey in this function.

These functions evolved from theoretical use case scenario development (Smith et al., in press) and matched the requirements for TramMate described above, resulting from in-situ future behaviour scenarios conducted by Kjeldskov et al. (2003).

Upon entering all required input, the system attempted to find a solution within the tram network between the origin and destination. The solution suggested by the system was optimal in terms of journey length (measured in number of stops), and the timing of tram vehicles.

The user interface and two screens from the *Plan Trip* option are shown in Figures 2 and 3 respectively.

Figure 2. Mobile route-planning service interface

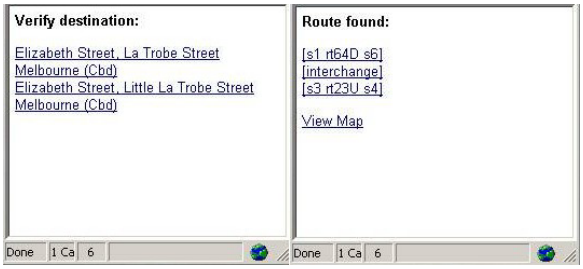

Figure 3. Sample screens from *Plan Trip*

5. Method

In order to conduct this study we studied five users in each of the laboratory and field studies, utilising a between-subjects design. The evaluation took place over three days with two days spent on field evaluations and one day being spent on the evaluation in the laboratory.

The output of each condition consisted of usability problems, task number, user, severity and supporting quotation. Both evaluations utilised the same series of tasks in order to make the user data from each evaluation comparable. The user had to complete these tasks using the mobile route-planning service. Examples of these tasks are shown below. These tasks were piloted for both the laboratory and the field in order to establish if they

were realistic and achievable within the time scheduled for each evaluation. The researchers wanted the tasks to cover the three main components of the system's functionality, be as realistic as possible and involve the user doing some work s/he was likely to perform with the service. From pilot studies it was established that the initial tasks should be reduced.

These tasks have been shown to have a high validity in terms of being relevant in users' lives and were judged to be realistic by the participants. As the system was a prototype, some of the functionality was limited. The tasks helped to cover the functioning of the whole system to enable its optimal range of use.

There were four main tasks taking the user through the three access points of the system. Their sequence constructed a logical (difficulty of task) and physical structure (movement through the city):

- Task 1: Utilisation of 'Timetable Lookup'
- Task 2: Utilisation of 'Plan Trip'
- Task 3: Utilisation of 'Determine route' (GPS)
- Task 4: User has free choice of 3 components

Tasks 2 and 4 are shown below.

Task 2: You want to catch a tram from the corner of Swanston and Queensberry Street in Carlton for a meeting at the corner of Little Collins and Exhibition Street in Melbourne. You have to be there about 30 minutes from now.

Using the "plan trip" option, find out:

a. Which tram route(s) to take
b. When the first possible tram is departing
c. The number of route changes (if any)
d. If there is a route change, where to board the second tram
e. Which stop to get off the last tram
f. How to get from the last stop to your final destination
g. The estimated time of arrival

Use this information to get to the meeting.

Task 4: You have finished eating. You are at Bourke Street Mall and want to return to the main entrance of Melbourne University

a. Use the system to get there as soon as possible

5.1. Users

Half the users were male and the other half were female, with male and female users balanced across the laboratory and field studies. Users were aged between 21 and 42, with a median age of 28 years. Nine users had completed an undergraduate degree with five having completed a Masters degree. Nine out of ten users were involved in academia and all were frequent computer users. All had used mobile devices and had some knowledge of the tram system and the Melbourne CBD.

All metadata were based on these 10 subjects. For the user data analysis, three users in each of the field and the lab study were examined closely due to time constraints. The six users were chosen in order to achieve as homogenous a sample as possible with regard to age and education. These users were aged between 21 and 25.

5.2. Field data

In the field, users firstly gathered the information required for the first two tasks while stationary and then performed the remainder of the tasks "for real". During this process the user was observed by three evaluators. One interacted with the user directly, encouraging the user to think aloud and asking questions in a manner similar to that used in contextual interview.

Figure 4. Field evaluation set-up

Another evaluator took notes on the user's interaction with the device, completing a data sheet with task number, time, notes/problems, important user comments and possible design improvements fields. These notes served as a backup in case the video did not capture enough data and were not used in the analysis. The third evaluator recorded the user, focusing on the user's interaction with the device and on capturing user comments. The user was encouraged to think aloud. The configuration for the field evaluation without interviewer is shown in Figure 4.

5.3. Laboratory data

For the laboratory a specialist facility was used which enabled the capture of the user's interaction with the device and the user's voice. A quad display captured the device screen, the user's face and the user's movement of the device. Typically, the evaluation focused on the user's interaction with the device's screen. The user was requested to use the device within a limited area indicated. The configuration of the laboratory is shown above (figure 5 and figure 6).

Figure 5. Laboratory evaluation: the participant

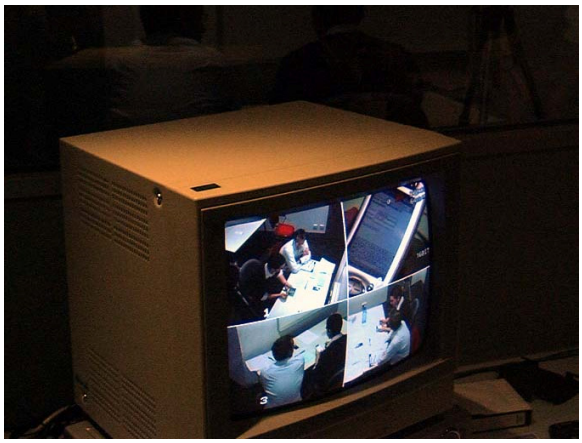

Figure 6. Laboratory evaluation set-up: observers´ point of view

5.4. Rapid reflection: metadata

Based on the two sets of collected user data, a rapid reflection process was used to collect metadata which was used for later comparison with the analysed video data. This metadata had three main components: diary entries from researches, minutes of meetings and observer's notes.

After each day of evaluation, the evaluators met to discuss the following questions:

- Have we reached a critical mass in terms of number of users?
- What are the main themes & application problems emerging from the data?
- What are your thoughts concerning this evaluation method?

The aim of these questions was to promote discussion of the usability problems emerging after each day's work. These questions were discussed and emergent themes were agreed upon, and comprised the meeting minutes. Diaries were kept by each of the evaluators over the three days of the evaluation. The observer's notes were collected by a researcher outside the data collection through observation of the evaluator meetings.

6. Analysis

Grounded techniques were chosen for the description of both the 'traditional' user data and the metadata for a number of reasons. Firstly, the user data collected was contained in multiple modalities. Grounded theory (Strauss & Corbin, 1998) possesses excellent tools for the analysis of such data. Secondly, the evaluators wanted to be able to compare the results of the two analyses easily: having similar representations for the final data sets made this possible.

Notably, the field and laboratory user data used the theoretical filter of usability problems to evolve descriptions of the data much earlier in the process of analysis. The metadata were analysed using a more grounded approach, where the user data were not approached with a particular theoretical understanding at the beginning of the analysis.

6.1. User data: field and laboratory

Usability problems were generated using qualitative analysis of the video captured in the field and in the laboratory. This analysis involved the joint analysis of one field user at the beginning of the analysis and one lab user at the end of the analysis to ensure that comparable results were generated. The video data were allocated to evaluators randomly. A matrix of usability problems, including task, time, problem number and description, user, severity, quotation, suggestions for improvements and additional comments was generated for each user. The matrices for the lab and the field were then summarised into two matrices through a grounded analysis by two of the researchers. This involved underlining keywords and repeated words in order to describe what main issues were in the memos. The evolved themes were cross-checked in order to triangulate the user data. To evolve the main categories or themes affinity diagrams were then used. For each emergent category the degree of severity was discussed. The severity was awarded according to Nielsen & Molich's (1990) definitions:

Critical problem:

- Recurred across all users
- Stopped users completing tasks

Severe problem:

- Recurred frequently across users
- Inhibited /slowed down users completing tasks
- Users could (eventually) complete tasks

Cosmetic problem:

- Did not recur frequently across users
- Did not inhibit users severely
- Users could complete tasks

The user data from the laboratory and the field was then compared with regard to the number and type of usability problems generated by each condition.

6.2. Metadata: field and laboratory

The metadata were also analysed using a more grounded approach. The emphasis of the analysis was on the meeting minutes. Keywords were identified in the metadata and evolved into themes through representation and description. Again affinity diagrams were used to evolve the main categories. For each emergent category the degree of severity was discussed. The severity was defined according to their frequency of occurrence in the minutes into major, prominent and minor. The data sets generated from the field and the laboratory study are represented in the results section below.

6.3. Comparison: data and metadata

In order to compare the two data sets (the metadata and user data), definitions of all themes were firstly checked and refined by both researchers separately. Then the metadata usability problems were compared to the usability problems collected from the field and in the laboratory. This was done by comparing categories, and checking and refining definitions across data sets. This comparison was conducted by both researchers independently and then results were compared and reconciled. In order to confirm that this comparison was defensible, the researchers then utilised the diaries recorded by the evaluators and an independent observer during the evaluation. The evaluators' diaries contained reflections on the process of conducting the field and laboratory studies and the independent observer's diary contained reflections on the meetings conducted. The results of these analyses are shown below. We do not report on the *cosmetic* themes from the laboratory and field user data and the *minor* themes from the metadata.

7. Results

7.1. Usability problems: user data

The following section describes themes with explanations and examples that emerged from the video analysis.

7.1.1. Field data

CRITICAL THEMES

System versus real world

This theme refers to issues caused by the relationship between the information contained in the system and the information contained in the world. Issues within this theme included street labelling being poor, tram stops not being marked on the maps, mapping between the system and the real world not being accurate and users not being clear about which tram to catch.

Maps

This theme refers to issues relating to system maps. Issues that emerged regarding maps concerned the lack of map clarity, the destination not being clearly marked on the maps, the user not knowing where to embark/disembark, there being no textual support or self-representation and the directions on the maps being poorly marked and not having stop numbers.

SEVERE THEMES

Input

This theme refers to issues related to the user inputting information into the system. This theme manifested itself when the user had a lot of information to input, did not know how to input information, became confused concerning the order of inputting information, did not know what to input and when the affordances were poor or when the PDA auto-complete function became cumbersome.

Prior knowledge

This theme relates to issues regarding the user's knowledge of computing systems and user knowledge of the environment in which they are interacting with the system. Two sub-themes emerged here: the user's need for knowledge of the city and tram stops and the lack of fit between the user's knowledge and the knowledge provided by the system.

Cognitive load

This theme relates to the load imposed on the user's memory and attention (especially memory) by the system. Specific issues under this theme included the user having to remember stop numbers and route information and there being no cognitive aids within the system.

User confidence

This theme relates to how confident the user is when using the system or confidence engendered by the system. Two themes emerged here concerning confidence in maps and the user not being confident about where to disembark.

Information

This theme relates to the relevance and accuracy of information contained in the system. Specific issues under this theme include missing information required by the user, the information not being tuned to context well enough or being too specific and part-whole relationships among information and distribution of information not being clear.

Navigation

This theme relates to how easy the user found it to move around the system screens. Three issues emerged under this theme: the user had to use too many clicks to do his/her work, the user could not go back easily and did not know where to go next

System issues

This theme related to the emergent features of the system dictated by use. This theme was manifested in the system not being flexible enough to adapt to user needs.

7.1.2. Laboratory data

CRITICAL THEMES

Information

This theme relates to how and what information is presented by the system at a certain time.

Issues under this theme included the distribution of information across screens being problematic, missing connectivity (especially between the maps and route details), inconsistency of the information on the screens and poor readability of information. This theme also described the relationship among the different screens not being clear to the user, the user expressing a desire for a different layout (particularly in tabular format) and the user not obtaining right information at the right time.

Navigation

This theme relates to how the user navigates through the screens of the system. The main issues emerging from this theme were the unclear structure of the screen connections and the user not knowing how to go back or where to go next in the system.

SEVERE THEMES

Input

This theme relates to the ease of input to the system and the affordances offered by the system. Specific issues under this theme included the user having difficulties inputting the required information via the virtual keyboard and not knowing what order to input information (first street corner and then suburb). In addition it was not clear what level of detail (e.g. if it was necessary to type in "street") was required when inputting information.

Prior Knowledge

This theme relates to how much knowledge of the city is required. The issue emerging from this theme relates to the user not being able to use the system without a considerable amount of city knowledge (such as the street corner and suburb of origin and final destination).

Cognitive load

This theme relates to the amount of cognitive resources needed to use the system. Two themes emerged here concerning the additional need of cognitive aids (paper) to write down the given information and the need to revise stop numbers.

User model

This theme relates to what the user's model of the system is. The issue emerging from this theme relates to the user model engendered by the system often not matching prior experiences, such as the use of landmarks or paper timetables.

Maps

This theme relates to how the user interprets and uses maps in conjunction with the textual information. The main issue under this theme is the lack of clarity caused by missing labels/route descriptions and unclear icons. The user also often did not know where s/he is located in the map nor in which direction to go and expressed the desire for a zoom function.

Labelling

This theme relates to how well the wording and symbols are understood by the user. Abbreviations like '*rt'* for route were not understood. Terms like *store* instead of *continue* and *earlier* and *later* without clear reference points did not make sense to users in context.

7.1.3. Summary: field & laboratory data

The user data are summarised in the table below:

Field	Laboratory
MAJOR THEMES (critical)	
System versus real world	Information
Maps	Navigation
PROMINENT THEMES (severe)	
Input	Input
Prior knowledge	Prior knowledge
Cognitive load	Cognitive load
User confidence	**User model**
Information	*Maps*
Navigation	*Labelling*
System issues	

Table 1. Usability problems from user data

Both methods reveal a high degree of overlap among themes (unique themes are in bold and themes that occurred in both studies and were only minor in one are in italics). The severity of problems in the different settings is interesting. It seems that classic usability problems, like the presentation of the information and navigation, are discovered by the traditional laboratory study and are very prominent, meaning these are severe problems. Looking at the same problems in the field, it is not that these problems do not appear as usability problems, but a shift of importance occurs between the empirical settings. While the laboratory problems connected to the system are extremely important, in the field the interaction with the device in the actual environment results in even more severe problems. These are more noticeable as problems in a rich context, over a laboratory setting. Obviously only the interaction with the system in the "real" world uncovers a mismatch of information in the system and information provided by the environment. This difference occurred between laboratory and field setting when participants were using maps and trying to match them with street corners and landmarks in the city. A similar difference between the laboratory and field data is revealed with the "user model" appearing in the laboratory, but not in the field. The participants tried to reflect on former experiences with real world information, such as with paper timetables or the use of landmarks in maps, because they did not possess real world information. A problem that was not discovered in the laboratory setting was "user confidence". This theme reflect social characteristics in our daily lives and is not a traditional usability problem

connected to a system. Further exploration is required to discover how a field experiment is able to explore these social components of using a mobile device.

7.2. Metadata themes

The following section only describes themes and explanations that evolved from the minutes of the rapid reflection during the debriefing sessions (meetings held among evaluators after each day of evaluation).

7.2.1. Field data

MAJOR THEMES

Interface adaptivity

This theme describes issues relating to the dynamism of the interface, information adaptivity and "contextness".

System versus real world

This theme refers to the dynamic between the system and the real world in terms of information and representations.

Input & constraints

This theme describes issues relating to difficulty of input and constraints on input imposed by the system.

PROMINENT THEMES

Directions

This theme describes data relating to the system not containing information on directions.

Information

This theme describes the amount and specificity of information presented by the system being a problem and part-whole relationships among the data.

Usefulness

This theme describes how the usefulness of the information is limited.

Efficiency

This theme describes the lack of efficiency users experienced using the system.

Dependency/trust

This theme describes issues related to the user's trust of and dependency on the system.

7.2.2. Laboratory data

MAJOR THEMES

Information

This is a very broad theme, including issues relating to the nature of the information being too specific, information being distributed across too many screens and the system's ability or inability to show relationships among general and specific route information.

Maps

This theme refers to lack of clarity in the maps, the relationships among maps not being clear and the maps having problematic symbolism.

PROMINENT THEMES

Navigation and information structure

This describes issues relating to moving around the system and the structure of information in the system.

User model

This theme refers to issues relating to and problems associated with how users thought the system should work.

Input

This theme describes issues relating to difficulty of input.

Symbolism

This theme describes issues and problems related to symbolism within the system such as problematic abbreviations.

7.2.3. Summary: field and laboratory metadata

The data are summarised in the table below:

Field	Laboratory
MAJOR THEMES	
Interface adaptivity	Information
System versus real world	**Maps**
Input & constraints	
PROMINENT THEMES	
Directions	**Navigation & information structure**
Information	**User model**
Usefulness	Input
Efficiency	*Symbolism*
Dependency/trust	

Table 2. Usability problems from metadata

There is a considerably lesser degree of overlap among themes across the two settings in the metadata compared to the user data (unique themes are in bold and themes that occurred in both studies and were only minor in one are in italics). Notably, in the field unique themes were related to emergent use (e.g. "usefulness"), whereas in the laboratory study unique themes related to attributes of the system (e.g. "maps"). In addition, the field metadata contain themes that are not covered by traditional usability problems (e.g. "dependency/trust"). The "user model" theme appearing in the laboratory and not the field again seems to reflect the lack of contextual information available to the user. The difference in the salience of themes such as "efficiency" again seems to reflect the ability of the field setting to test real use. In this regard, the laboratory setting seems more able to identify specific interface issues.

7.3. Comparison of data sets

A direct comparison of the two data sets leads to the results shown in Table 3. This comparison is limited to the major and prominent themes. In the field user data "system versus real world" appears in the field metadata as major a theme as well. "Input" and "information" are

both themes appearing in the field user data and metadata. Apart from this all other themes are labelled differently. However, there is still an overlap as regarding the content of the themes as evidenced by the above explanations.

In the two sets of laboratory data the major and prominent themes overlap to a very high degree. For example the major theme of "information" in the user data is also shown by the metadata to be crucial. There is an even higher similarity when looking at the prominent themes. However, the themes "cognitive load" and "prior knowledge" that are important themes in the user data cannot be found in the metadata. The "labelling" theme appeared as a sub theme within the laboratory metadata in the form of "symbolism".

Method	*User data*	*Metadata*
Field	SIMILAR	
	System versus real world	
	Input	
	Information	
	DIFFERENT	
	Maps	Interface adaptivity
	User confidence	Dependency/trust
	System issues	Efficiency
	Navigation	Directions
	Cognitive load	Usefulness
	Prior knowledge	
Laboratory	SIMILAR	
	Information	
	Navigation	
	Maps	
	User Model	
	Input	
	DIFFERENT	
	Labelling	Symbolism
	Prior knowledge	
	Cognitive load	

Table 3. Comparison: user data and metadata

In sum the metadata themes are on a more abstract level. The user data provide more concrete examples while the metadata are on a more general level like the category "usefulness". Categories like "indexicality" and "personal user strategies" (minor themes that are not listed here) seem to reflect researchers' early subjective theories. On the other hand, themes that were only sub-themes in the user data, like "symbolism", become prominent within the metadata. The metadata emerged from an overall impression of different users while the user data categories were developed user by user. So recency effect, primacy effect and possible other forms of data biases are more likely to occur in the metadata.

7.4. Metadata: diaries

Additional findings concerning the evaluations emerged from a grounded analysis of the diary entries and observer notes. The most notable findings concerned the field evaluation, with evaluators describing the following issues:

- the effort involved to conduct the study and if the effort involved was worthwhile;
- the validity and generalisability of the data generated from the study;
- the ability to capture the richness of the user's context through the approach.

Over time confidence in the method grew and contentment with the results increased, with one evaluator commenting it generated "rich and useful data". Main worries concerned practical problems like:

- capturing the data in an appropriate way;
- influencing the user with early hypotheses and suggestive questions;
- interpreting the data in an objective way (researcher reliability);
- the lack of a theory of mobility describing dynamic use contexts.

This metadata was particularly useful when reflecting on the effectiveness of the methods used in this study.

8. Discussion and conclusion

The analysis of the data and metadata created important results regarding the use of the different methods, as well as the use of a combination of user data and resulting metadata. In relation to the secondary objective described in section 2 the field study discovers relevant problems for mobile use that are not covered by traditional sets of usability problems. Mobile use involves much more than the interaction of the user with the device. Rich and dynamic use contexts have to be accounted for in mobile evaluation methods. The strength of the laboratory study lies in a clear and easy discovery of usability problems. Advantages of one or the other methods depend on the fidelity of the prototype and aim of the study. In the field study this level of fidelity was just sufficient. A lower degree of fidelity would probably have lead to problems for the user and would have made an evaluation difficult. Thus the choice of the method is closely related to the fidelity of the prototype. If the focus of the study is on usability problems and an amendment of the system itself, it is reasonable to start with a laboratory study.

The user data show that a field study is valuable for mobile devices. However, the method still needs improvement particularly regarding data capture. The following citation from the diary of one of the evaluators summarises one of the problems:

"*The sun meant that video capture was difficult. Even at the beginning, focus on the 'rich' data was hard: screen detail was often drowned by background noise. My feeling was that video would only provide context and would have little value during analysis*"

In addition, a better definition of the method itself is needed. Being a mixture of traditional and a more

constructivist approach the quality of the data is difficult to judge. Again, a quote from the diary of one of the researchers illustrates:

"*My feeling was that less focused work for the users would have been better. We were neither doing an observation nor doing a usability test: it seemed a strange hybrid of methods at times.*"

Regarding the primary objective the cost efficiency of the analysis of the field and laboratory user data and metadata, the following can be concluded. The combination of a laboratory study and an analysis of metadata at an early state of the prototype seems useful. If the aim is to determine the usefulness of the device, a field study is highly recommended. Here the collection and analysis of field metadata is still useful, but is not sufficient to discover the rich usability problems emerging from mobile use. The metadata support the user data set in a useful way but do not replace the systematic analysis of the user data.

Metadata give a quick and helpful insight into the main problems. Overall, the metadata seem to have a higher reliability for the laboratory data than the field data. Reasons for this might be the missing dynamic use context that produces many impressions in the evaluators. Additionally the participation as an evaluator in the laboratory setting seems less straining and tiring. Another contributing factor may have been that the evaluators involved in this study were more experienced with conducting laboratory evaluations and therefore could draw conclusions more effectively from observing users in that settings, thereby generating metadata that reflected the user data better.

Quality criteria like internal and external validity demand a certain number of participants to produce statistical power. The comparison of metadata and data raises the question of how these criteria can be meaningfully discussed. Having a mixed approach involving traditional research and constructivist/ naturalist paradigms can be a problem. Further research has to be done to define appropriate criteria in a useful way and to reconcile these different paradigms. The concept of mobility necessitates new methods and therefore new ways to measure and judge them.

In this context it is important for future research to reflect how to map methods for evaluating mobile devices clearly to certain process steps. The methods used depend highly on the fidelity of prototype, novelty of the product to the user, time and other resources. Further studies are needed at which stage of the design process it is useful to include collection of metadata or even use exclusively metadata as an analysis technique.

9. Acknowledgements

Thanks to colleagues on the TramMate project: Steve Howard, Jennie Carroll, Daniel Tobin, Frank Vetere and John Murphy as well as the Novell employees who participated in the field studies. Special thanks are reserved for Steve Howard for his valuable input into the paper and for coining the phrase "rapid reflection".

10. References

Braiterman. J. & Larvie, P. (2002). Each sold separately: ethnography as a tool for integrating online and off line use of educational toys. *Proceedings of the HF2002 Human Factors Conference*. Melbourne, Australia

Gray, W. D. & Salzman, M. C. (1998). Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods. *Human Computer Interaction 13,* 203-261

Guba, E.G.,& Lincoln, Y.S. (1989). Fourth generation evaluation . Newbury Park, CA: Sage.

Henderson, R., Smith, M., Podd, J., & Varela–Alvarez, H. (1995). A comparison of the four prominent user-based methods for evaluating the usability of computer software. *Ergonomics 39*, 2030–2044.

Karat, C-M, Campbell, R L, Fiegel, T (1992), "Comparison of empirical testing and walkthrough methods in user interface evaluation", in P. Bauersfeld, J. Bennett, & G.Lynch (Eds.), *Proceedings of CHI'92. New York: ACM*, pp. 397-404.

Kjeldskov , J., Howard, S., Murphy, J., Carroll, J. Vetere F. and Graham C. (2003). Designing TramMate-a context aware mobile system supporting use of public transportation. *Proceedings of Designing User experiences – DUX 2003,* San Francisco, ACM

Kjeldskov, J. & Graham, C., 2003. A Review of Mobile HCI Research Methods, Mobile Human-Computer Interaction: *Proceedings of the Mobile HCI 2003 Conference. Udine, Italy*

Kjeldskov, J. & Skov, M. (2003). Evaluating the Usability of Mobile Systems: Exploring Different Laboratory Approaches. *Proceedings of the HCI International 2003*, *Crete, Greece* Vol. 2, p.122-127

Luff & Heath (1998). Mobility in collaboration. Computer Supported Cooperative Work archive. *Proceedings of the 1998 ACM conference on Computer supported cooperative work table of contents. Seattle, United States* p. 305 – 314

Millen D. R.: Rapid Ethnography (2000): Time Depending Strategies for HCI Field Research. Symposium on Designing Interactive Systems 2000: 280-286

Nayak, N. P. Mrazek D., Smith R.D. (1995). Analysing and Communicating Usability Data. *SIGCHI Bulletin 27 (1)*

Nielsen, J. and Molich, R. (1990). Heuristic evaluation of user interfaces. *Proc. CHI'90 Conference on Human Factors in Computer Systems. New York: ACM, 1990*, pp. 249-256.

Olson, G. M., Moran, T. P. (1998). Introduction to this Special Issue on Experimental Comparisons of Usability Evaluation Methods. Human Computer Interaction V13, pp.199-201: Lawrence Erlbaum Associates, Inc.

Smith, J., Mackaness. W., Kealy, A. and Williamson, I.P., (in press), Spatial Data Infrastructure Requirements for Mobile Location Based Journey Planning. *Transactions in GIS*.

Strauss, A. L. & Corbin,. (1998). Grounded Theory in Practice. Sage Publications