Mobile Information Systems II

Edited by
John Krogstie
Karlheinz Kautz
David Allen

Springer

ifip

MOBILE INFORMATION SYSTEMS II

IFIP – The International Federation for Information Processing

IFIP was founded in 1960 under the auspices of UNESCO, following the First World Computer Congress held in Paris the previous year. An umbrella organization for societies working in information processing, IFIP's aim is two-fold: to support information processing within its member countries and to encourage technology transfer to developing nations. As its mission statement clearly states,

> *IFIP's mission is to be the leading, truly international, apolitical organization which encourages and assists in the development, exploitation and application of information technology for the benefit of all people.*

IFIP is a non-profitmaking organization, run almost solely by 2500 volunteers. It operates through a number of technical committees, which organize events and publications. IFIP's events range from an international congress to local seminars, but the most important are:

- The IFIP World Computer Congress, held every second year;
- Open conferences;
- Working conferences.

The flagship event is the IFIP World Computer Congress, at which both invited and contributed papers are presented. Contributed papers are rigorously refereed and the rejection rate is high.

As with the Congress, participation in the open conferences is open to all and papers may be invited or submitted. Again, submitted papers are stringently refereed.

The working conferences are structured differently. They are usually run by a working group and attendance is small and by invitation only. Their purpose is to create an atmosphere conducive to innovation and development. Refereeing is less rigorous and papers are subjected to extensive group discussion.

Publications arising from IFIP events vary. The papers presented at the IFIP World Computer Congress and at open conferences are published as conference proceedings, while the results of the working conferences are often published as collections of selected and edited papers.

Any national society whose primary activity is in information may apply to become a full member of IFIP, although full membership is restricted to one society per country. Full members are entitled to vote at the annual General Assembly, National societies preferring a less committed involvement may apply for associate or corresponding membership. Associate members enjoy the same benefits as full members, but without voting rights. Corresponding members are not represented in IFIP bodies. Affiliated membership is open to non-national societies, and individual and honorary membership schemes are also offered.

MOBILE INFORMATION SYSTEMS II

IFIP International Working Conference on Mobile Information Systems, (MOBIS)
Leeds, UK, December 6-7, 2005

Edited by

John Krogstie
Norwegian Institute of Science and Technology and SINTEF
Norway

Karlheinz Kautz
Copenhagen Business School
Denmark

David Allen
Leeds University, School of Business Administration
United Kingdom

Library of Congress Cataloging-in-Publication Data

A C.I.P. Catalogue record for this book is available from the Library of Congress.

Mobile Information Systems II
Edited by John Krogstie, Karlheinz Kautz, and David Allen

p. cm. (IFIP International Federation for Information Processing, a Springer Series in Computer Science)

ISSN: 1571-5736 / 1861-2288 (Internet)
ISBN-10: 0-387-29551-8
ISBN-13: 9780-387-29551-0
Printed on acid-free paper

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 11571841 (HC)
springeronline.com

Contents

Preface

The second IFIP TC8 working conference on mobile information systems (MOBIS) was held in Leeds 5-6 December 2005. The objective of the working conference was to provide a forum for researchers and practitioners across the IFIP TC8 working groups interested in planning, analysis, design, construction, modification, implementation, utilization, evaluation, and management of mobile information systems to meet, and exchange research ideas and results. Specifically, the working conference looked at:

- the adaptation and organizational impact of mobile information systems
- existing and newly developed approaches for analysis, design, implementation, and evolution of mobile information systems
- technical issues and the constraints they impose on mobile information systems functionalities and design.

The conference would not have been possible without the assistance of many people. We received 38 papers which were reviewed by a minimum of two reviewers and 16 full papers and 7 short papers were accepted. We are indebted to the program committee members and additional reviewers for preparing thorough reviews on a very tight schedule. We like to thank the authors for their efforts to make excellent scientific contributions to this new and challenging field. Finally, IFIP, TC8 officers and the local organizers have been instrumental for the success of the event

September 2005

John Krogstie, Trondheim
Karlheinz Kautz, Copenhagen
David Allen, Leeds

Conference Committe

Program Co-Chairs

Prof. John Krogstie, IDI, NTNU and SINTEF, Norway
Prof. Karlheinz Kautz, Copenhagen Business School, Denmark

Organizing Chair

Dr. David Allen, Leeds University Businees School, UK

Program Committee

Kalle Lyytinen, USA
Barbara Pernici, Italy
Keng Siau, USA
Guttorm Sindre, Norway
Mikael B. Skov, Denmark
Kari Smolander, Finland
Anthony Wasserman, USA
Robert Steele, Australia
Marie Thilliez, France
Elaine Lawrence, Australia
David Simplot, France
Carl Adams, UK
Jan Damsgaard, Denmark
Jennifer Blechar, Norway
Dewald Roode, South Africa
Siggi Reich, Austria
Richard Baskerville, USA
Jari Veijalainen, Finland
Matti Rossi, Finland
Jo Herstad, Norway
Erik Gøsta Nilsson, Norway
Luciano Baresi, Italy
Boualem Benatallah, Australia
Chiara Francalanci, Italy
Manfred Hauswirth, Germany
Wolfgang Prinz, Germay
Jens Wehrmann, Germany
Gordana Culjak, Australia
Lim Ee Peng, Singapore
Binshan Lin, USA
George Giaglis, Greece
Sherif Kamel, Egypt
Hannu Kangassalo, Finland
Carsten Sørensen, UK
Pamela Coutts, Australia
Claudio Bettini, Italy

TAKE RISKS INTO CONSIDERATION WHILE JOB DISPATCHING

Shi-Cho Cha, Hung-Wen Tung, Chih-Hao Hsu, Han-Chao Lee, Tse-Ming Tsai and Raymund Lin
Advanced e-Commerce Institute, Institute for Information Industry, Taipei, Taiwan

Abstract: To deal with the uncertainty of job dispatching in mobile workforce management, we have proposed a framework, *Risk-Oriented joB dispAtching for mobiLe workfOrce system* (ROBALO) (Cha et al., 2005), to ease the tension between (a) the reliability requirement to serve a job request, and (b) the cost of the job's assignment. In ROBALO, the risks for workers to execute a job are taken into consideration. Such consideration is especially useful in the scenario of mobile workforce management because mobile workers usually meet unexpected situations in the field. Therefore, we can find the job assignment with the minimum cost under a certain degree of risk. Therefore, the job dispatcher can reserve enough resources and make enough preparations for a incident. Our previous work focuses on the scenario of online job dispatching, which chooses a worker or a working group to serve every incoming job request independently. In this article, we further extend to the batch model and propose a sub-optimal approach for batch job dispatching to form a complete framework.

Keywords: Mobile Workforce Management System, Online Job Dispatching, Batch Job Dispatching, Risk Management

1. INTRODUCTION

The framework of ROBALO, abbreviated term of *Risk-Oriented joB dispAtching for mobiLe workfOrce system* (Cha et al., 2005), is proposed to handle the uncertainty of job dispatching in mobile workforce management. In traditional job dispatching mechanism, exception handling processes are usually taken as the only counter-measure in dealing with the failure of job execution. Compared with this approach, ROBALO takes risks into consideration. Therefore, the time to discover the failure can be saved because we try to do things right at the first time. Furthermore, while mobile workers may usually meet the unexpected situations in the field, ROBALO would be very useful in the scenario of Mobile Workforce Management Systems (MWMSs), for it considers the uncertainty in the real world.

Simply speaking, ROBALO is made to enable a systematic approach to ease the tension between (a) the reliability required to fulfill a specific request and (b) the cost of the assignment. This mechanism would make it capable to find the assignment with the minimal cost in a certain degree of risk, so that the dilemma of reliability and cost could be balanced. Furthermore, the availability of risk information would make it possible for the job dispatcher to reserve resources and make some preparations for the exception.

In (Cha et al., 2005), it discussed the scenario of online job dispatching. Online job dispatching chooses a worker or a working group to serve every incoming job request independently, which is useful for emergency calls. When time is available, the batch model can be used to assign workers to finish several known jobs; in such case, although the cost for a specified job may increase, the overall cost may be decreased. In this article, we extend to the batch job dispatching schemes and show how ROBALO can be implemented to take risk consideration into job dispatching, so as to achieve the benefits mentioned previously.

The rest of this paper is organized as follows. Section 2 introduces the framework of ROBALO. Section 3 shows how ROBALO measures risks for a worker or a working group to do a certain job. Section 4 and Section 5 discuss how risks can be used for online and batch job dispatching respectively. Section 6 surveys related work on job dispatching in mobile workforce management. Finally, conclusions and future work are offered in Section 7.

2. OVERVIEW OF ROBALO

The architecture of ROBALO is depicted in Figure 1. First of all, the related information of job dispatching can be provided in the following:

- The *Scheduler Manager* manages the *schedules of workers*.
- The *Profile Manager* maintains the *profiles of workforce*, the *workforce's current statuses*, and the *job execution historical logs*. The profile of workforce contains workers' basic information, capacities, and other demographic information. The workforce's current statuses include workers' current location, availability information, and etc., which can be tracked by the *status monitor*. (But, it is beyond the scope of this paper to discuss status monitoring.) The results of *job execution historical logs* are recorded for a worker or a working group to execute a job.
- The *Cost Evaluator* maintains the cost for a person to execute a job in a *cost matrix* and other auxiliary information, such as the traffic cost information, to calculate the cost for a service provided by the worker.
- Finally, the *Case Manager* maintains the profiles of each case and the historical logs of job execution failure.

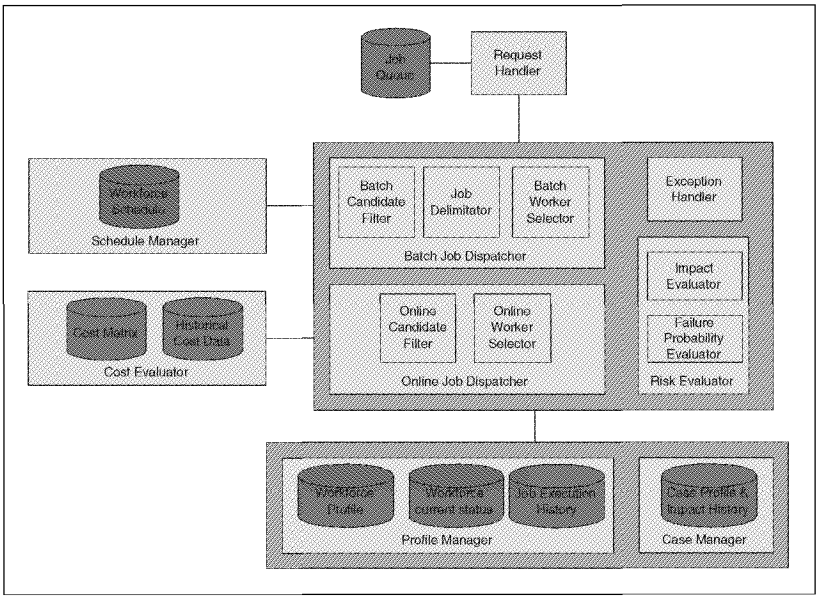


Figure 1. ROBALO Architecture

When a service request is received, the *Request Handler* parses the requests to extract the context information and generate preferences and constraints of the service. Provided that the request is required to be processed immediately, it is firstly sent to the *Online Job Dispatcher*. The *Online Candidate Filter* is in charge of filtering out the workers that are not capable of doing this work and obtains a candidate list for this job. The *Risk Evaluator* predicts the risks for a worker or working group in charge of the job in the candidate list, while the risks are calculated on the basis of the predicted loss and probability of failure, which will be further discussed in details in Section 3. Then, the *Online Worker Selector* selects the workers or working groups to serve the job request on the basis of their risks and costs. In the end, the job will be dispatched to the selected workers.

If requests do not need to be processed online, they can be kept in the *job queue*. ROBALO is used to group the requests into batches by time periods, e.g. 24 hours, and to send the various jobs to the *Batch Job Dispatcher* by batch. For example, in a job dispatching center, the jobs are assigned on the daily basis at 10:00pm for the next day. The time of a job batch can be further divided by *Job Delimitator* into several sub-periods. In each sub-period, a worker can only be assigned to do one job at this period. With such simplification, local optimization solution in each sub-period can be found. The

details are as shown in Section 5. The Batch Job Dispatcher executes its jobs dispatching sub-period by sub-period as follows: At first, similar with the Online Candidate Filter, the *Batch Candidate Filter* excludes out the workers who are not qualified to do any of the jobs in the sub-period. The Risk Evaluator predicts the risks of a job to be assigned to each worker. The Cost Evaluator estimates the cost for a worker to fulfill the task assigned including the job execution, the travel costs, and etc. The *Batch Job Selector* then tries to find out the minimized cost solution with the risks constraints.

Finally, there are few more factors to be taken into account, such as some jobs which no qualified workers could be found, in either online or batch job dispatching process, or in the case that even though the risks have been looked after, the execution of the job might still be failed. At this phase, the exceptions could be dealt with *Exception Handler*. However, the details of exception handling are beyond the scope of this article.

3. RISK PREDICTION

This section show how risks are estimated. In literatures, the definition of risk is extraordinarily broad and inconsistent. Some definitions adopt the traditional view of risks - the possibility or potential impact of unwanted occurrences. Other definitions include the uncertainties inherent in achieving optimal performance, success or failure in seizing available opportunities, and etc... (COSO, 2004). Herein we adopt the traditional definitions of risk (because it is widely used in the domain of information security (Tipton and Krause, 2004)) and calculate risks of a job assignment to be the multiple from the probability of execution failure with losses arising.

In remainder of this Section, Section 3.1 shows how the impact is anticipated. Section 3.2 demonstrates the estimation process of failure probability. In the Section 3.3 it illustrates how risks can be calculated from related losses and probabilities.

3.1 Impact Evaluation

Figure 2 shows the input and output of the Impact Evaluator. First of all, the context of a request can be defined as follows:

DEFINITION 1 (CONTEXT) *The context of a request r (we denote it as C_r) is represented by a m-ary tuple: (c_1,c_2,....., c_m). Each c_i represents an attribute or a feature of the context.*

With the definition of context, we can further define the impact history:

DEFINITION 2 (IMPACT HISTORY) *The impact history H can be defined as a set of tuple (C_i, I_i). For a past failure request R_i, C_i is the context of*

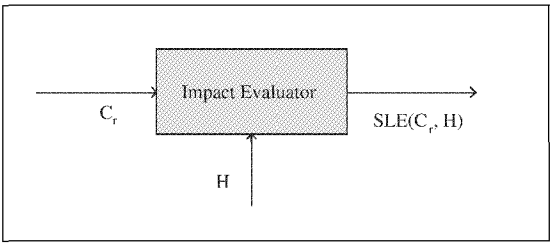


Figure 2. The I/O of the Impact Evaluator

this request. And I_i is the loss incurred from the failure execution of the job request.

The output of the Impact Evaluator is the prediction of the loss when the request r cannot be finished on time (it is denoted as $SLE(C_r, H)$). The traditional linear discriminant analysis methodology (Fisher, 1936),DK1982 can be used as follows:

- Firstly, we classify the amount of loss into several classes. For example, we can divide the amount of loss into five classes as Table 1. For each value in $Class_i$ and $Class_j$, the value in $Class_i$ is less than the value in $Class_j$ if i is less than j.
- Secondly, for each adjacent class, find the linear discriminant function $g_{i,i+1}$ with the impact history H, where $g_{i,i+1}(X) = \sum_{1 \leq k \leq m}(v_k \times x_k + v_0)$ and X is a m-ary tuple: (x_1,x_2,....., x_m). These linear discriminant functions are re-calculated in a certain interval so that these functions can reflect the up-to-date scenarios. For a tuple X, if $g_{i,i+1}(X) \leq 0$, X belongs to $Class_k$ where $k \leq i$. Otherwise, X belongs to $Class_l$ where $l > i$.
- Finally, when a context C_r is received, we calculate which class C_r belongs to as shown in Figure 3. And the ceiling of the range of the class is returned as $SLE(C_r, H)$.

3.2 Probability Evaluation

As shown in Figure 4, the Probability Evaluator predicts the probability that a working unit fails to finish a job under the context of the received request. Besides context, we have the following definitions:

DEFINITION 3 (WORKING UNIT) *A working unit W_r is the minimum set that can be assigned to serve a request r. Suppose that the universal set of workers is U ($U = \{u_i \mid u_i$ is a worker$\}$), $W_r \in U*$.*

Table 1. Classification of Impacts

Class	*Range*
$Class_1$	\$0 – \$10K
$Class_2$	\$10K – \$1M
$Class_3$	\$1M – \$100M
$Class_4$	\$100M – \$10G
$Class_5$	\$10G –

1.*m = the number of classes*
2.*for i=1, i≤m-1*
3. *if* $g_{i,i+1}(C_R) \leq 0$ *then* C_R *belongs to* $class_i$
4.*next i*
5.C_R *belongs to* $class_m$

Figure 3. The Algorithm for Context Classification.

DEFINITION 4 (JOB EXECUTION HISTORIES) *The job execution history EH can be defined as a set of tuple (C_i, W_i, S_i). For a past job request R_i, C_i is the context of this request, W_i is the work unit assigned to execute the job, and S_i is its result (either success or failure).*

DEFINITION 5 (PROFILES OF WORKFORCE) *The current profiles of workforce WP is represented by a set of x-ary tuple: (wp_{i1},wp_{i2},....., wp_{ix}). Each wp_{ij} represents an attribute or a feature of the worker u_i's profile.*

DEFINITION 6 (WORKFORCE STATUSES) *Similar to the profiles of workforce, the current statuses of workers WS is represented by a set of y-ary tuple: (ws_{i1},ws_{i2},....., ws_{iy}). Each ws_{ij} represents a kind of status of the worker u_i.*

To make it simple, we assume that the failure probabilities for each worker to execute a job is independent on another. This means that the failure probability of a work unit $P(W_r, C_r, EH, WP, WS)$ can be calculated from the product of $P(\{u_k\}, C_r, EH, WP, WS)$ where $u_k \in W_r$. Before computing $P(\{u_k\}, C_r, EH, WP, WS)$, a linear discriminant function d is found as follows:

- For each element (C_i, W_i, S_i) in EH, we extract the members of the working unit.
- For each worker u_k in the working unit W_i, we obtain the worker's profiles WP_{kt} and statuses WS_{kt} at that point of time t.

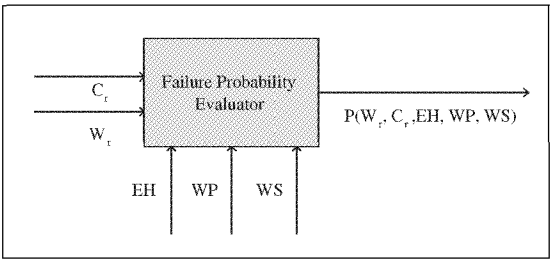


Figure 4. The I/O of the Probability Evaluator

- We concatenate WP_{kt}, WS_{kt}, and C_i into a new $(x+y+m)$-ary tuple (we call it as a working features vector (wf_{kti})). And use the tuple to find out the linear discriminant function d.
- After the linear discriminant function d is found, we calculate wf_{kti} for each working features vector and generate a set D. In the set D, we have tuples of the value of $d(wf_{kti})$ and its related result, e.g., (0.12, success), (-1, failure), (0.01, failure), etc..... And we denote it as (d_{wf_i}, s_{wf_i})

Then, $P(\{u_k\}, C_r, EH, WP, WS)$ can be calculated as follows:

- First of all, we extract the user's profiles and statuses and concatenate them with C_r into a working features vector wf_k.
- Compute the value of $d(wf_k)$.
- Compare the value of each d_{wf_i} in D and find N-nearest tuples.
- Compute the number of tuples where the value of s_{wf_i} is failure. If the number is n, the probability of failure is predicted as n/N.

Finally, $P(W_r, C_r, EH, WP, WS)$ can be obtained from the product of $P(\{u_k\}, C_r, EH, WP, WS)$ (for each $u_k \in W_r$).

3.3 Risk Calculation

DEFINITION 7 (RISK) *If we select a working group $S(S = \{u_i | \, u_i \, is \, a \, worker\})$ to execute a job, the risk of the job assignment R_r is the product of its loss expectancy ($SLE(C_r, H)$)) and its failure probability ($P_r(S)$).*

We assume that the failure probability for a worker to execute a job is independent to others. Therefore, we can calculate $P_r(S)$ from the product of $P(W_i, C_r, EH, WP, WS)$, for each $W_i \in S$. That is, $R_r = SLE(C_r, H) \times \prod_{W_i \in S} P(W_i, C_r, EH, WP, WS)$.

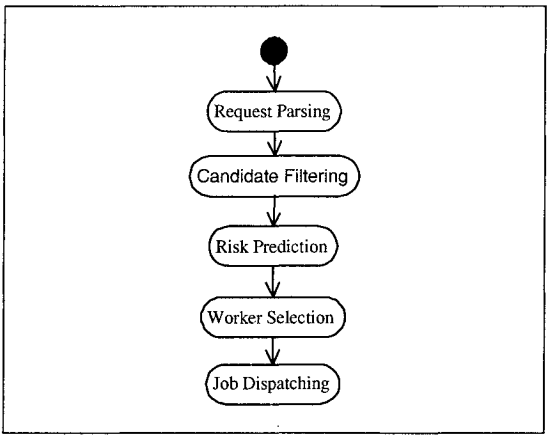


Figure 5. The Process of Online Job Dispatching in ROBALO

4. ONLINE JOB DISPATCHING

Figure 5 shows the process of online job dispatching. As described above, when a job request is received, the Request Handler parses the request to extract the context information from the request and generate constraints of the job. The Candidate Filter then uses these constraints to filter out unqualified workers and obtains a candidate list for this job.

At this point, we say a worker is capable of doing a job if he or she can satisfy the requirement of the job. For a job, its requirement includes:

- *Capability requirement.* For each job, there may be several requirement about workers' capability. For example, the worker must have a specified license or something.

- *Status requirement.* We denote the status required to execute a job J_i as SC_i (SC_i is a y-ary tuple of (sc_{i1},sc_{i2}, ...,sc_{iy})). For example, the status tuple could be a unary tuple of a location where the location is the place to execute the job. The statuses of a worker u_i is WS_{it} at time t (WS_{it} is a y-ary tuple (ws_{it1},ws_{it2},ws_{it3}, ..., ws_{ity})). In this case, a worker u_i is capable of doing a job J_j if he can change his/her status to satisfy J_j's status requirement. That is, ws_{it1} satisfies sc_{i1}, ws_{it2} satisfies sc_{i2}, etc.....

- *Time requirement.* For each job J_i, it must be executed between JS_i and JE_i. That is, a worker must be able to change its status to satisfy the status requirement of J_i before JS_i and finish the job before JE_i.

After the request context and candidate list are obtained, for each worker in the list, ROBALO evaluates his/her risk to finish the job by the product of the predicted loss and failure probability as mentioned in Section 3. And the Cost Evaluator calculates the cost for the worker to execute the job. Suppose that the cost for a worker u_i to be assigned to do a job J_j is C_{ij}. C_{ij} can be calculated from the sum of the cost (CW_{ij}) of job execution and the workers' travelling cost (CS_{ij}). The former can be obtained from a cost matrix. And the latter can be predicted by estimating the cost for a person to change his/her statuses to satisfy the status requirement before the job's time requirement JS_j.

Suppose that the maximum acceptable risk is R_a. With the candidate list L_c, the risks for a worker to execute the job, and its cost, we can select S with the pseudo-code shown in Figure 6. Generally speaking, the algorithm wishes to find S with the minimized cost where $P_r(S) \times SLE(C_r, H) \leq R_a$ or $P_r(S) \times SLE(C_r, H) \leq R_a/SLE(C_r, H)$. Obviously, if $R_a/SLE(C_r, H)$ is greater or equal to 1, any job assignment can satisfy the constraint. Therefore, we can select the worker W_i in L_c with the least cost. Otherwise, because $0 \leq P(W_i, C_r, EH, WP, WS) \leq 1$, we can apply the logarithmic function to the inequality $\Pi_{W_i \in S} P(W_i, C_r, EH, WP, WS) \leq R_a/SLE(C_r, H)$. The line 6–line 19 in Figure 6 try to find a solution to $\Sigma_{W_i \in S} log(P(W_i, C_r, EH, WP, WS)) \geq (log(R_a) - log(SLE(C_r, H)))$ with the minimized cost. The greedy algorithm is used here to choose the worker with the biggest unit cost of $log(P(W_i, C_r, EH, WP, WS))$.

5. BATCH JOB DISPATCHING

In Figure 7, it shows the process of batch job dispatching. In addition to parsing requests, differ from online job dispatching, the Request Handler collect the requests that do not need to be processed online and keeps them in its job queue. It then sends a batch of job requests required to be executed in a period of time. For a job J_i, suppose that the job is required to be executed between JS_i and JE_i, it is said that a job is required to be executed in a period of time if JS_i is within this period. Suppose that the beginning and the end of a duration is d_s and d_e, for each job J_i in d ($d_s \leq JS_i \leq d_e$), we cannot find another job J_j where $d_s \leq JS_j \leq d_e$, $JS_i < JS_j$. Based on the delimitation, we can reduce the dispatching problem to assign n single jobs to m workers.

Figure 8 gives an example about job delimitation. There are four jobs in a period P. J_0 does not belong to this period because it starts before the start time of P. We first find the job with the least termination time among the jobs started in this period. Then, we use the termination time of this job as the end of the first duration. Take Figure8 for example, at first, we find that J_1 is the job with the least termination time in P. Therefore, we can obtain the first

1. *If* $R_a/SLE(C_r, H) \geq 1$ then
2. *Select the worker* W_i *in* L_c *with the least cost*
3. S=$\{W_i\}$
4. *return* S
5. *Else*
6. *current_risk = 0*
7. $W' = L_c$
8. $S = \{\}$
9. *while current_risk*$>R_a$ *and* W' *is not empty*
10. *Find the worker* W_i *in* W' *with the biggest* $logP(W_i, C_r, EH, WP, WS)/C_{W_i}$
11. $W' = W' - W_i$
12. $S = S \cup W_i$
13. *current_risk* = $SLE(C_r, H) \times \Pi_{W_i \in S} P(W_i, C_r, EH, WP, WS)$
14. *End While*
15. *If current_risk*$>R_a$ *and* W' *is empty then*
16. *The result cannot be found*
17. *return* ϕ
18. *End If*
19. *return* S
20. *End If*

Figure 6. The Pseudo-code for Worker Selection.

duration D_0 where the start time of D_0 is the start time of P and the end time of D_0 is the termination time of J_1. Then, we find J_3, which is the job with the least termination among the jobs in the remaining period. Therefore, the duration D_1 can be defined by using the end time of D_0 as the start time of D_1 and the termination time of J_3 as the end time of D_1. Similarly, we can further divide the remaining period into durations until there is no jobs starting at the remaining period.

After dividing the period into durations, we dispatch jobs duration by duration. To dispatch jobs in a duration, the Batch Candidate Filter filters out the workers who are not capable of doing any jobs in this duration. It is the similar case of Online Candidate Filter mentioned in Section 4 except that it takes several works into consideration at the same time.

At this point, we have a candidate list of workers and jobs needed to be dispatched in the duration. ROBALO then estimates the risk and cost for a worker in the candidate list to be assigned to do a job. We predict risk in the same way as online job dispatching. However, we need to make some modifications while estimating the cost. In online dispatching, we use a worker's current statuses to estimate the cost for the worker going to do a job by predicting the cost of the worker to change from their current statuses to the statuses satisfied with the requirement of the job. However, in batch job dispatching, a worker's statuses may not be known unless his or her previous schedule is determined. For example, if a worker is assigned to do a job in place A from

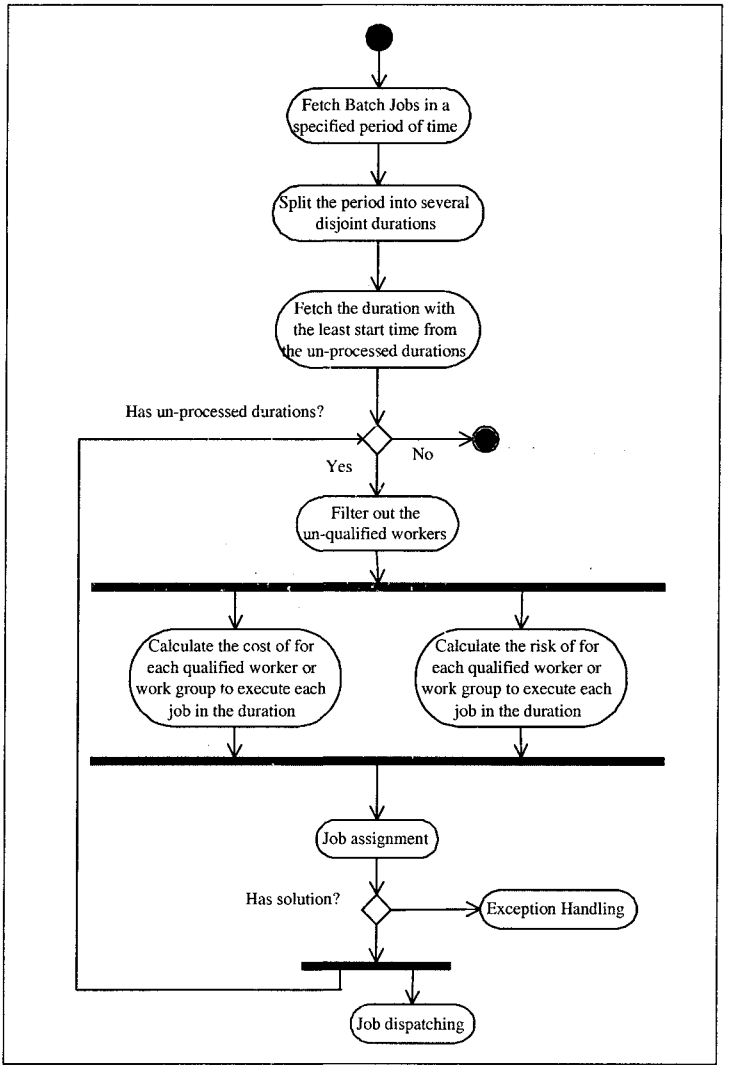


Figure 7. The Process of Batch Job Dispatching in ROBALO

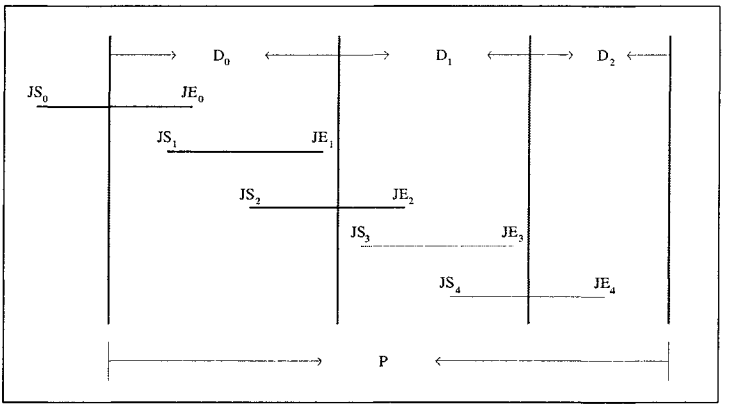


Figure 8. A Example of Job Delimitation

place B in the afternoon, the travelling cost can be reduced if the worker is assigned to do another job in place A in the morning.

Because we do job assignment from the earliest duration, when we wish to estimate the cost of changing a worker's statuses to acceptable statuses for a job, we can find the nearest previous statuses of the worker. Therefore, we can estimate the cost of changing the worker's statuses in that point to the acceptable statuses of the job and use it to calculate the cost for the worker to be assigned to do a job.

Before assigning jobs to capable workers, the Batch Worker Selector scans the risk (R_{ij}) for a worker u_i to do a job J_j. If we find that the risk is higher than a predefined acceptable level of risk R_a ($R_{ij} > R_a$), the Batch Worker Selector set related cost as infinite to present the job to be assigned to the worker.

Then, the Batch Worker Selector can simply assign workers to the jobs by solving the following linear programming problem:

Minimize $\Sigma_{u_i \in L_c, J_j \in J} C_{ij}$ (C_{ij} is the cost for worker u_i to do job J_j. L_c is the set of qualified candidates), subject to:

$\Sigma_{u_i \in L_c} x_{ij} = 1$

$x_{ij} = 1$ *or* 0 (x_{ij}=1 if J_j is assigned to u_i)

Based on the above procedure, we can make sure that we can find a local optimization solution with the minimized costs for batch job dispatching.

6. RELATED WORK

Traditional job dispatching methodologies can be classified into the following two categories: (1) the batch model and (2) the online model.

The batch model assumes that the job requests and available workers are known in advance. Then the available workers are assigned to deal with these given jobs. The batch model is useful for assigning the routine or scheduled jobs. However, in some cases, we cannot know every job requests before. For example, an emergency center may receive an emergency call from a injured person and need to dispatch an ambulance to take him/her to a nearby hospital. At this point, we can use online model to assign one or more workers to serve the incoming request.

Current online job dispatching schemes usually assign workers to serve a incoming job request based on the workers' cost, capabilities, or current statuses. For example, HAMS (Healthcare Alert Management System)(Chiu et al., 2004), (Aydin et al., 2004) classifies staffs in a hospital into different roles based on their capacities. And each kind of tasks (or alerts) can only be assigned to the staffs with appropriate roles. When a alert is triggered, HAMS select a person from the available staff members who can play the roles required for the alert and dispatches the alert to him/her. The SOS Alarm (Normark, 2002) dispatch ambulances for emergency calls based on the proximity to the ambulance stations and the status of each ambulance.

Besides workers' cost, capacities and status, we propose to take the probability that a person may fail to finish a job into consideration and use risks as a factor for job assignment in this article. While we evaluate the risk before job assignment, we can predict the expected loss of the assignment and make some "preparation" (e.g., we can assign another person as backup) while the job is dispatching. Traditional job dispatching systems usually initiate their exception handling processes after it finds that assigned workers fail to finish a job. For example, HAMS (Chiu et al., 2004) and (Aydin et al., 2004) sends messages to all staffs with the same role as it finds that a staff does not confirm a job assignment. At this point, in comparison with traditional exception handling approach, the time to discover the failure can be saved because ROBALO tries to do things right at the first time.

Moreover, risk also provides a systematic way to balance the consideration between reliability and cost consideration. This is because ROBALO assigns just enough workers so that a job can be finished within a specified level of risk. Therefore, resources can be used in a efficient way because a job dispatcher can get rid of assigning many workers to do a job blindly.

7. CONCLUSION AND FUTURE WORK

In this article, the implementation issues about ROBALO are described in details. Generally speaking, ROBALO takes the risks for a worker or a working group failing to execute a job assignment into consideration. Such consideration is especially useful in the scenario for mobile workforce management

because mobile workers may usually meet unexpected situations in the field. With ROABLO, the tension can be eased between (a) the reliability requirement to serve a job request, and (b) the cost of the job's assignment by finding the assignment with the minimum cost under a certain degree of risk. Therefore, job dispatcher can reserve enough resources and make enough preparation for an incident. In traditional job dispatching mechanism, the exception handling processes are taken to deal with the failure of job execution. In comparison with this approach, time of failure discovery can be saved because the rights things are done at the first time.

Other than a concrete implementation of ROBALO, there are many interesting things left to be done. First of all, in the current design of ROBALO, it assumes that the failure events are independent, while the relationship between workers is taken into considered. Secondly, in batch job dispatching, a period is divided into several sub-periods so as to find the local optimal solution with linear programming schemes. Moreover, it is assumed that a job can only be assigned to one worker for simplicity. It is rather a complicated however interesting problem worthy for further studies to find an overall optimal solution. Finally, although, in the ROBALO system, linear discriminant programming analysis is used to predict the risks from history data, more complicated approaches, such as non-linear programming technologies, could be included.

ACKNOWLEDGEMENTS

This research was supported by the Service Web Technology Research Project of Institute for Information Industry and sponsored by MOEA , ROC

References

Aydin, N., Marvasti, F., and Markus, H. S. (2004). Embolic doppler ultrasound signal detection using discrete wavelet transform. *IEEE Trans. on IT in Biomedicine*, 8(2):182–190.

Cha, S.-C., Tung, H.-W., Lee, H.-C., Tsa, T.-M., and Lin, R. (2005). Robalo: A risk-oriented job dispatching mechanism for workforce management system. In *IFIP I3E2005*.

Chiu, D. K. W., Kwok, B. W. C., Wong, R. L. S., Cheung, S.-C., and Kafeza, E.(2004). Alert-driven e-service management. In *HICSS*.

COSO (2004). *Enterprise Risk Management – Integrated Framework*. COSO.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annal of Eugenics*, 7:179–188.

Normark, M. (2002). Sense-making of an emergency call: possibilities and constraints of a computerized case file. In *Proceedings of NordiCHi'02*, pages 81–90, New York, NY, USA. ACM Press.

Tipton, H. F. and Krause, M. (2004). *Information security management handbook 5-th ed.* CRC Press. ISBN 0-8493-1997-8.

SUPPORT OF SMART WORK PROCESSES IN CONTEXT RICH ENVIRONMENTS

Carl-Fredrik Sørensen, Alf Inge Wang, and Reidar Conradi
Dept. of Computer and Information Science, Norwegian University of Science and Technology (NTNU), N-7491 Trondheim, Norway.

Abstract: The evolution of mobile and ubiquitous technologies gives promises for computational services and resources to support and influence work processes planned or performed in physical work environments. Such support should optimally provide the workers with a safer work environment for both the environment itself and the workers. The extended support can give more economic and optimal work processes through proactive and situated planning and execution. We introduce the concept of a *smart work process* to capture adaptive and context-aware process support. This combination of ubiquitous computing and workflow defines a new research direction to be investigated. This paper elaborates on research challenges related to how smart work processes can be supported. We present and discuss general cases where context information or change in context information influences mobile work activities. Finally, we propose a framework for modelling smart work processes, and present a high-level architecture to support smart work processes.

Keywords: Mobile work, Mobile and ubiquitous environments, Smart work processes

1. INTRODUCTION

The future mobile and ubiquitous computing environments hold promises for computational services and resources to become invisible but important parts of the supporting computing environment for all kinds of user activities. It can be used to improve the accessibility to information and computational services. The information sources can be very diverse, from simple sensors sampling environmental properties, to complete information

systems with the ability to roam e.g. the Internet for information and services. Mobility imposes a dynamic and unstable environment that challenges how to create mobile work support environments. New opportunities to be utilised by the mobile workforce include management of ad hoc activities and cooperation with other people, as well as exchange of information and services with the surrounding environment.

Today, mobile computing and communication support is mostly directed to provide availability of services while being mobile through wireless networks and mobile computing devices. We consider this support to be an extension of existing distributed and wired services to make it possible to work distributed at different places using mobile computers.

The demanded computational support will differ from user to user, and from activity to activity. The needs are based on, e.g., what process or context information available or required by the user, and under which conditions an activity is to be performed. The conditions are influenced by temporal changes of context information as well as the behaviour of humans or other actors in the environment. Together, they constitute the state of the mobile working environment.

In this paper, we discuss and differentiate mobile work from nomadic work, and then identify issues related to how different types of context information can be supported and utilised in systems to support mobile work processes. We therefore introduce the notion of the *smart work process*. We further propose a process framework and architecture to support smart work processes.

2. MOTIVATION

Many work places have a complex structure of participants, activities, and artefacts. This makes even simple work processes hard to plan and execute. Such environments are dynamic and unpredictable. Resources can be scarce and different participants can therefore compete to make necessary adaptation in the environment to fit their own goals. Often in such environments, safety is a very important issue for the employers and other actors in the society.

Safety for single actors can be ensured by establishing safe working conditions to perform activities within, or to re-establish safe conditions by performing situated or planned activities. Actors can influence the working environment in different ways that can create safety-breaks for other actors as well as "stealing" resources from others. By providing context information to the individual and to supervisory process enactment services, it is possible to initiate coordination activities to establish safer and more

economic working conditions. **Smart work processes** can be used as a means to coordinate multiple actors. Reactiveness to environmental changes is thus important to ensure an optimal safety and production rate. In environments that are self-aware, i.e., equipped with augmented, "intelligent" artefacts (Strohbach, 2004), activities can be initiated by the environment to ensure certain environmental goals. In addition to provide activities, coordination of actors and sequencing of activities can be performed to maintain or enhance productivity, safety, or other goals defined in the environment, by the actors or by their processes.

3. MOBILE WORK ENVIRONMENT

Mobile work has often been looked upon as an extension of distributed work in terms of technological solutions to support such processes. Support of distributed work includes systems to manage configuration of shared artefacts, and to manage coordination and collaboration among the participants. Mobile work support systems may have need for these properties as well, but mobile work is clearly distinct from distributed work by the motivation to be *mobile* and how the dynamic change of physical environment influences how to perform work. The supporting infrastructure suffers from unpredictability and availability to the people that work in a mobile setting. In addition, mobile workers also have to face dynamic and partly unpredictable changes in the physical environment. To establish a clear definition of mobile work, *mobility* must get more emphasis to distinguish mobile work processes from distributed, co-located or individual work processes. Mobility must be a property necessary to perform the actual work, irrespective of technology. This means that work that is independent of mobility, cannot be characterised as mobile work, even if the work is performed when mobile, i.e., change of location is not a pre-condition for performing the work. We can therefore split work processes performed in a mobile setting into two different categories based on context-sensitivity:

- **Work in a mobile environment**: Work processes performed in a mobile environment *independent* of context information extracted from the physical environment. That is, mobility is *not necessary* to accomplish the process goals.
- **Mobile work**: Work processes performed in a mobile environment *dependent* of context information extracted from the physical environment. That is, mobility is *necessary* to accomplish the process goals.

A simple public transportation scenario can illustrate the difference: A bus is transporting people from one place to another. The driver is

performing mobile work because he has to adapt to changes in the physical environment. A bus passenger working with a laptop connected to a wireless network is not performing mobile work but work in a mobile environment. Figure 1 illustrates the differences between nomadic work, service work, and inherent mobile work. **Nomadic work** refers to **anywhere, anytime computing**, (often called *nomadic computing*). The goal is to provide users with access to popular desktop applications, applications specially suited for mobile users, and basic communication services in a mobile, sometimes wireless environment (La Porta, 1996). Such work is not regarded as *context-dependent*. **Service work** refers to workers that need to travel to a specific location to perform work. The work is thus context-dependent, but the context can be regarded as quite stable. **Inherent mobile work** requires continuously change of location and thus a dynamic environment. All work performed when in a mobile situation may require similar technological and computational support.

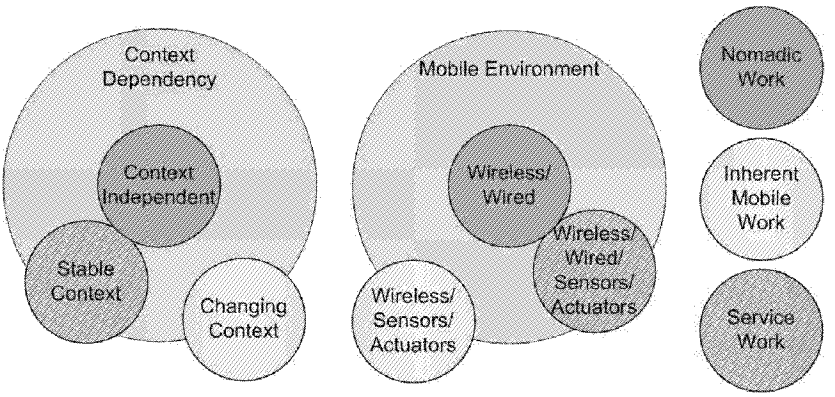


Figure 1. Mobile Work vs. Working when Mobile

To support mobile work, the evolution in mobile computing technology enables access to distributed workflow systems. However, the mobile environment is dynamic, and processes depending on changes in properties of the local environment are not particularly supported by such systems. Thus, for mobile work, it is necessary to have access and support related to the local environmental context. Traditional distributed workflow system with enabling mobile technology is therefore not sufficient to support mobile work processes.

Mobile work and the local working environment can mutually influence each other in different ways:

- *Mobile work can change the state of its environment.* This can be done directly by initiating and performing activities with the purpose of

changing the environmental state, e.g., by introducing actuators or other equipment to control or change the environment. The environmental state can also be changed indirectly through activities performed by one or more actors (workers) that together introduce state changes.

- *The state or change of state in the environment* can influence many different aspects of how, when, by whom, and what to be done in an environment. The environmental state is thus directly or indirectly used in the planning and enactment of the work process. The state or state change can:
  - *Initiate activities that must be performed.*
  - *Decide start, duration, delay, stop, and termination of activities.*
  - *Decide which activities to be performed.*
  - *Change the content or goal of an activity.*
  - *Initiate exceptions in the current activities.*

Activities and change of state in the environment together provide dynamics to reach process goals by adjusting each other. This point relates to how an environment can be regarded as an organism with a certain amount of order (state) infused with chaos through activities, where the order and chaos must balance to successfully reach a pre-defined or situated process goal.

We will in next section go into deeper details how to support mobile work in a mobile working environment.

4. SMART WORK PROCESSES

In this section, we unify mobile work with environmental context-awareness to propose a framework for process models to be used to support mobile work. To be able to sense the environment, we need methods and tools to capture relevant context information that affects the mobile work process. This information will be necessary to drive the work support application; we therefore denote all externally sensed information as *context.*

To provide the required process support to deal with work processes that influence or are influenced by the mobile environment, we introduce the notion of a **smart work process**:

*A work process that is **sentient** (i.e., sense the environment which it performs in), **adapts** to relevant context or context changes through **context-based reasoning** to reach process goals, and **actuates** by providing situated activities, or by changing/refining the current planned activities.*

Smart work processes are thus especially tailored to make use of context information to monitor and coordinate activities within a context-rich environment. Importantly, coordination is the glue for how to manage

singular activities within an environment. Coordination needs to be performed between specific work processes and the work environment, between multiple actors performing possibly cooperative or competing work processes, and with respect to some stated paramount requirements like safety, time, and economy. The contextual relationships are not directly related to the normal relationships that exist between activities within ordinary, office-like work processes or plans. There is therefore a need for extracting and specifying how work support applications can be adjusted or adapted to also cover smart work processes.

4.1 Context-Aware Activities

An activity can be defined in a process model by providing goals, pre- and post-conditions, invariants, and the use or production of artefacts and resources. The process model elements may be affected by context information in various ways. Context or context changes can in such a model be used as pre-conditions whether to start, stop, or terminate an activity. The activity is thus directly affected by and dependent on the contextual state and hence need to include specifications and rules for how the context types are related to the activity. Context can also provide rules for how and what to do in an activity. In this case, the context is used to specify activity content and thus refining and concretising the abstract definition of an activity. Further, context can provide alternative process paths as described in Section 3. The process paths can either be pre-specified or need in situ specification when encountering new states not covered by the model. Context can also be used to trigger or create new activities not previously defined in a process model. This case is related to situated actions (Suchman, 1987) and can possibly be handled through situated planning as described in (Bardram, 1997). This means that the process model needs to be extendable in runtime to cover in situ specification and enactment.

Physical work or an actuation resulting from an activity can directly or indirectly change the physical environment (post-condition). Such a change can either be predictable and enforced through the activity and thus can be specified, or be a side-effect. Side-effects must either be included into the process model in cases where such effects are preferable, or be met by reciprocal activities to counter the side-effect. In any case, side-effects must be handled by the process enactment service. In some cases the environment is in such a state that it is necessary to perform activities to provide pre-conditions to perform new or (pre- or situated) planned activities. A related case is activities to prepare the environment for a certain planned activity. This requires a process model to be extendable based on how activities are dependent on certain environmental states to be executed.

The cases above identify how a process enactment service must take into account the surroundings to effectively support the enactment of smart work processes. The process model is in addition to the changes in the process plan and goals, also affected by state changes of elements in both the physical and the computational environments. The elements in the environment may be affected by the process, but also by other factors that might be out of control from the enactment service point of view. The environment may contain other "competing" actors, artefacts, resources and other elements that cause coordination problems between actors, the physical and computational environment itself, and the process goals of the different actors.

4.2 Situated actions and planning

The state of the physical work environment will often be fluctuating and dynamic, and it is thus very hard to provide realistic or detailed plans beforehand for which activities to be performed. Definition or refining of activities based on the perceived state of the environment will therefore be necessary to create executable process models. Some activities can also be defined in situ by the environment itself. This is possible if the environment is self-aware and augmented with applications and technology making the environment able to change its own state as well as ask for concrete actions by external or passing (human) actors not statically bound to the environment. The state of the environment provides means to create activities to change the environmental state to a level where a planned goal or intention may be accomplished.

Situated planning and thereby actions will therefore be a natural part of a support environment for smart work processes. Context information is thus used for both definition and adaptation of work processes in situ.

4.3 Challenges in context-aware process support

The support for context-aware or smart work processes is to a small degree covered in the literature and is thus a new area of research within pervasive computing and workflow/process support. Since the domain has little coverage in the research community, many challenges arise in the cross-section of mobile/wireless computing, ubiquitous/pervasive computing, and workflow:

- Specification of ***contextual pre/post-conditions*** related to some process goal.
- Specification of ***environmental behaviour*** related to an activity (adaptation).

- Specification of a ***uniform representation of sensors and actuators*** from a process enactment perspective to make it possible to reason about and change the context state of the environment.
- ***Planning, specification, and execution*** of activities in concert with the current environmental context. These challenges relate to how the dynamics are handled by the workflow enactment services, both on client and on server.
- ***Managing process changes to ensure a consistent state*** of the process.
- ***Managing the dynamics of ad hoc activities and process changes*** locally and in a central enactment service.

5. A PROCESS FRAMEWORK FOR SMART WORK PROCESSES

Process models are used to define how work should be performed and can be used for visualising processes, guiding the user through the process, automating steps of the process, educating users about new processes, analysing or simulating the process etc. To define a process model, a process modelling language (PML) must be used.

The Workflow Management Coalition (Hollingsworth, 1995) provides a reference model for how to build a workflow management system and how to make interfaces to other applications and workflow systems. To support smart work processes, some additional or changed requirements are necessary to provide extended support of context-aware work processes. We therefore specify some of the requirements to a process support system supporting smart work processes:

The process model must be able to ***adapt*** during the process enactment. This is partly covered through the use of exceptions in the reference model.

The process model must be ***refinable/extendable*** by situated planning. Working in a dynamic environment with coarse plans is quite normal. Such plans are refined or extended based on the properties of the working environment itself. Activities resulting from situated planning are not possible to plan beforehand, but systems can use encountered processes to "learn" about specific activities performed during specific situated work situations.

A ***separate environmental or world model*** must be built up to support specific processes by identifying relevant context sources that may affect or support the process.

Rules must be specified or developed for how the process model can adapt to the relevant context information. The rule model defines how the process and the environment may or should interact during the process

enactment. The rules may be pre-defined, but must also be derived from the need for situated actions, the sensed state of the environment, and the process itself. In addition, the environment can provide rules based on artefacts augmented with smart sensors and some representation of "intelligence" (Strohbach, 2004).

The working environment must have "services" that can be given values of certain required properties or conditions to be satisfied during the process or activity execution, e.g., the environment is supposed to keep a certain temperature during a process. A break in such conditions can contradict the process goal and in the worst case endanger the environment itself or the people in the environment.

A *coordination service* is necessary to coordinate the different actors with cooperating or competing process goals or plans. The actors might be loosely coupled through the environment and not directly aware of each other presence in the environment.

Specific environments can have pre-defined activities or plans that can be put into enactment in certain pre-defined environmental states, e.g., defined safety procedures when the safety requirements are not satisfied.

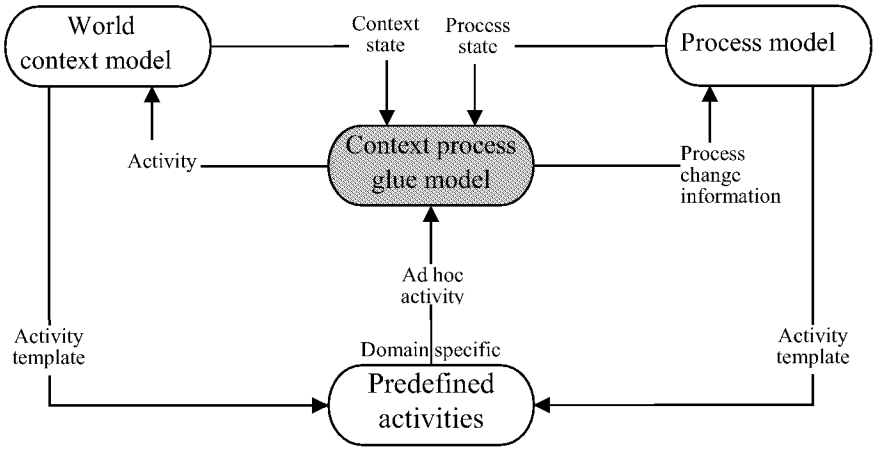


Figure 2. Smart work process glue model

Figure 2 shows a proposal of model needed to provide support for smart work processes and how these models relate. A central part of modelling smart work processes is the representation of the context relevant for the work process.

The *world context model* represents objects in a ubiquitous computing environment, often embedded with sensors and actuators. Such a world model should consist of all relevant static objects in an environment that somehow can influence the work processes. In addition, mobile objects that can provide relevant context information must also be integrated/included in

such a model. Also non-tangible properties like e.g. weather, sounds, and light conditions may be relevant and therefore need to be included in the model. The construction of a relevant world model poses challenges since this means that the world model itself must be able to change when objects are approaching or leaving the environment. The static world model is, however, easier to comprehend and therefore more easily represented in a model. Other challenges are how to specify and select the relevant objects in an environment to be included in the specification and enactment of context-aware work processes. Such a selection is necessary both when pre-planning the work process and when using situated planning. The world context model must represent the relevant objects in a way that make it possible to catch the state of the physical or execution environment. The state is needed to reason about and take according actions in the work process support system. A challenge here is to have a context model that is able to provide a unified representation of objects/sensors that can represent various physical data that can span from boolean values to complex, composite data types representing abstract properties in the environment. A *process model* is also needed that model the activities of the work process. Such a model must allow late binding and it must be possible to reconfigure and change the model during the process enactment.

A key in modelling and supporting smart work processes is the *context process glue model*. This model is a loose coupling between the world context model and the process model that defines rules for how the context should affect the process model and vice versa. The glue model is responsible to connect pre-planned processes with situated planning, and thus process enactment. For pre-planned processes that are context-sensitive, we proposed to use templates that can be refined during context-aware, situated planning. To deal with unexpected conditions in specific work environments, we propose to have a set of *predefined activities* used to handle such conditions. The predefined activities can dynamically be added to the process model and activated based on the new environmental state. A library of predefined activities will grow over time based on gained experiences on solving unexpected events. The predefined activities are influenced by the process model as well as the world context model.

6. AN INFRASTRUCTURE TO SUPPORT SMART WORK PROCESSES

The future computing environment is predicted to realise the vision of the invisible computer (Weiser, 2001). Computing devices are in this scenario embedded in almost every kind of physical object. Sensors are

spread like "dust" in the environment or appear as "brilliant rocks" (Satyanarayanan, 2002). This vision makes the computing environment extremely distributed and gives indications of an enormous amount of possible context sources and services to autonomously support people in their whereabouts. Augmented, intelligent artefacts can provide self-aware computing elements in the environment that can cooperate to provide rules and actions used by the process support system. In addition, actuators can provide services that can change the state of the environment. Such actuators can be robot-like and thus have the possibility to perform a limited number of activities (partly autonomous, mobile units with communication capabilities).

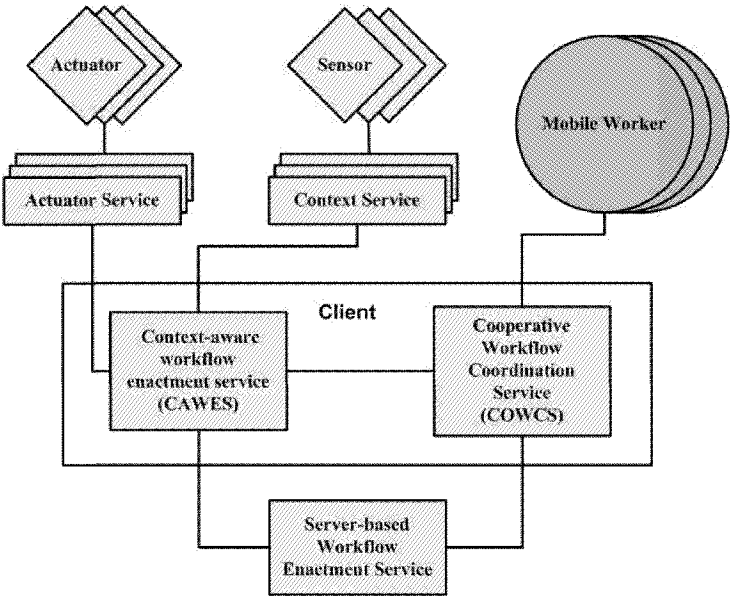


Figure 3. An Architecture for smart work processes

We have developed a high-level architecture as shown in Figure 3. The architecture consists of seven main parts which will be described in more detail below.

Actuator: An actuator is an entity that is able to change the environment based on digital input from some digitial equipment. In our architecture, an actuator is responsible for acting according to specifications instructed by the actuator service.

Sensor: A sensor is responsible for measuring, and transmitting sensor readings according to specifications by the context service. Sensors might vary from very simple, to autonomous, augmented entities.

Actuator Service: An actuator service is responsible to initiate actuations in an environment. The service should be able to receive instructions from the workflow enactment service and translate these to the actual actuator. The actuators can in this service be represented as abtract entities with published properties. Actuator services can hide how to communicate with the different actuators and provide a well-defined interface to be used by the service clients. The actuator service should be able to communicate with actuators through various physical communication implementations like IR, WLAN, LAN, BlueTooth, etc.

Context Service: A context service is responsible for identifying sensors, initiate sensor readings, check sensors, setup/terminate sensor subscriptions, etc., i.e., provide the communication with all actual sensors. The context service can accept subscriptions from different clients and is then responsible for setting up the sensor subscription properties. The clients can specify which properties they prefer to receive. The context service is responsible for handling client sensor subscription information including receiving rules, preferences, and facts related to sensor properties; collect/aggregate basic readings into abstract context (e.g. weather conditions). Based on the properties received from the various clients, the context service should send context messages to the client subscribers related to subscriber preferences. A context service may be extended to also represent e.g. augmented artefacts. It should then be able to communicate process fragments based on templates initiated by certain contextual states/conditions either contained within the context service, or received from smart, autonomous sensors. Possible context services that can be provided are e.g. weather conditions, engine and application states, etc. (including the state og the client itself), location, and the mobility of the user and other objects/entities within the environment.

Client-Based, Context-Aware Workflow Enactment Service (CAWES): CAWES is responsible for monitoring and executing delegated activity (ies), through surveillence of the current process/activity state, and contextual events received either from the context service, the COWCS (see below), or the Server-based Workflow Enactment Service. CAWES can also receive process/activity fragments from the context service, the COWCS, and the Server-based Workflow Enactment Service, can send its own process/activity state as contextual events to the COWCS, and the Server-based Workflow Enactment Service. If it is a need for actuations, CAWES initiates actuations using the Actuator service. Further, it can display/inform about environmental state wrt. the process itself, safety state, etc. It can also display/inform about information related to the activity performed. Such information can be based on the activity itself (manuals, checklists, or multimedia information), location-based information, safety information, etc. In additions, CAWES should also provide a traditional workflow user interface for filling in information, create/update artefacts related to the

activity, e.g. interactive checklists (can also be filled automatically using the context service), manage/integrate audio/video/textual information collected through the device or an actuator service), etc. An extension of CAWES is to provide virtual/augmented, location-based information (manuals, multimedia) related to the working environment, process type, available tools and sensors.

Client-Based Cooperative Workflow Coordination Service (COWCS): COWCS is responsible for managing and coordinating activities in a multi-actor environment. This includes a coordination service that based on policies can send messages to all involved parties about coordination needs, competitive resources, etc. COWCS receives process goals, and currently performed activities from the other actors. Similarly, it can send its current state of executing activity and possibly prepare other users about new activities to be started within a time limit, and coordination messages to other users including needs for artefacts, services, actuators, and a preferred environmental state. Internally on the client, COWCS sends contextual events to the CAWES based on coordination reasoning. Such events may include information about the need for artefacts, services, actuators and an environmental state with priorities, time, demand, etc. It can also send new process fragments derived from the coordination service or received from other clients based on demand/need from other users. The fragments are sent both to the CAWES, and to the Server-based Workflow Enactment Service.

Server-based Workflow Enactment Service: The Server-based Workflow Enactment Service is responsible for traditional workflow management including sending activities to the users, based on delegation/plans, and managing the complete, high-level process for the involved participants. The architecture has not been completely validated, but we have developed some prototypes to start the process of validating the architecture and the concepts behind it. Unfortunately, few work environments have been sufficiently instrumented to try out the proposed technology in practice. This situation, we hope will be improved in a few years.

7. RELATED WORK

(Wiberg, 1999) state that "anytime, anywhere" does not necessarily mean "everytime, everywhere". The ideal mobile situation is not to work continually without any stops. Further, true mobility goes beyond mobile support for "here and now". There is also a need to support the place to go, and the place left behind as well as to make plans for the future or backtracking earlier events. Mobile work is in many cases a kind of stationary work because the worker has to stop to perform any real work when work is physically oriented.

Situated actions (Suchman, 1987) and situated planning (Bardram, 1997) are terms that relate to our approach. It is hard to make predictions of which actions to perform in dynamic environments, i.e. the environment itself is influencing the actions based on the state of the environment. This motivates for a broader use of contextual information to support both the execution f actions, and also the planning of them. Situated planning can be used to refine coarse-grained plans based on the contextual properties of the environment and the worker itself. Situated actions can be used to create immediate plans to be used in situ.

The CORTEX project proposes the sentient object (SO) programming model (Fitzpatrick, 2002) for pervasive and ad hoc computing applications. Smart work processes can be looked upon as sentient objects. A smart work process can be decomposed into smart activities or process fragments that also can be regarded as sentient objects.

All the work presented above is important to enable sound architectures and designs for smart work processes. Thus, both conceptual and technical issues can be addressed by using similar design frameworks to create a workable infrastructure to support mobile work where context information plays an important role for a successful enactment of smart work processes.

8. DISCUSSION

We have in this paper not discussed all challenges that may arise when working in mobile, ubiquitous computing environments. We have, however, identified a few issues that may be hard to address to ensure a safe and dependable support of smart work processes:

Supporting process state transitions and process enactment when disconnected from a central work process support server: When disconnected we need to establish local enactment services in addition to a central enactment and coordination service. It must also be possible to coordinate the clients decentralised in an ad hoc fashion.

Ad hoc activities based on environmental context: The definition of ad hoc activities can be done in at least three ways: Such activities can be automatically defined by inferring activities through context-sensing and reasoning, they must be specified semi-automatically (using partly pre-planned templates), or be specified manually. This requires knowledge about the relationship between the context state and activities to be performed to reach an either pre-planned, or a situated goal. The activity must also be directed to the correct worker and then be coordinated with other activities in the environment. The state initiating the ad hoc activity can at the same time also influence the other activities in the environment creating conflicting

goals or incompatible activities. This requires coordination activities to solve inconsistencies as well as optimising the activity throughput in the environment.

Ad hoc collaboration between environment and users, user to user: Collaboration between the different actors in a working environment is often not specified explicitly. In some cases, the collaboration is specified in the PML, but often collaboration can be initiated by the current situation or state in the environment. It is a challenge to create support systems that are able to create such process ad hoc relationships.

Human considerations: Context-aware systems (Bellotti, 2001) must consider properties like intelligibility and accountability to give dependable systems providing added value to the user. We envisage the need to provide intermediate evidence of the applicability and dependability of the solutions, and that the number of different kind of context sources to be included in the systems must be slowly increased to keep the systems manageable and accountable. The importance of *which* context information to include *when,* will therefore be a paramount issue to evolve applicable rules. Learning processes should therefore be included to evolve the inference engines embedded in the systems.

Technical issues and solutions related to mobility, mobile devices, wireless networks, and the inherent properties of these, have not been covered in this paper. Such issues are important to address to successfully implement support for smart work processes.

9. CONCLUSION

We believe that the concept of *a smart work process* is useful as an abstraction of adaptive, context-aware work processes used to model and support mobile work in future ubiquitous computing environments. We have in this paper clearly distinguished mobile work from nomadic work to better illustrate how future process support should be developed. Current process modelling languages do not support situated processes directly influenced or created by the environment. Such processes are mainly unknown to the work process support systems until the situation has occurred in which the process is to be performed in. Context information captured in the working environment and used in work support applications can in the future be important to keep control of dynamic work environments. Coordination and implicit cooperation between the environment and participating actors can ensure a safer and more economic work environment directing and coordinating work activities using all relevant context information that can be captured using sensors and reasoning techniques. A lot of work remains

before the vision of smart process support systems can be realised. Especially in multi-actor environments and environments with autonomous equipment like robots, the need for supervision and adaptation of work processes is important to ensure safety, and economy through better utilisation of the available resources.

References

Atluri, V. and Chun, S. Handling Dynamic Changes in Decentralized Workflow Execution Environments. In *Lecture Notes in Computer Science*, pages 813–825. Springer-Verlag, 2003. LNCS 2736.

Bardram. J.E. Plans as Situated Action: An Activity Theory Approach to Workflow Systems. In *5th European Conference on Computer Supported Cooperative Work*, Lancaster University, UK, 7-11 September 1997. Kluwer Academic Publishers.

Bellotti, V. and Edwards, K.. Intelligibility and Accountability: Human Considerations in Context-Aware Systems. *Human-Computer Interaction (HCI) Journal. Special Issue: Context-Aware Computing*, 16(2–4):193–212, 2001.

Fitzpatrick, A, Biegel, G., Clarke, S., and Cahill, V.. Towards a Sentient Object Model. In *Workshop on Engineering Context-Aware Object Oriented Systems and Environments (ECOOSE)*, Seattle, WA, USA, November 2002.

Hollingsworth, D.. The Workflow Management Coalition – The Workflow Reference Model. Technical report, Workflow Management Coalition, Lighthouse Point, Fla., January 1995. Technical Report WFMC-TC00-1003, version 1.1.

La Porta, T. F., Sabnani, K. K., and Gitlin, R. D.. Challenges for Nomadic Computing: Mobility Management and Wireless Communications. *Mobile Networks and Applications*, 1(1):3–16, 1996.

Satyanarayanan, M.. Of Smart Dust and Brilliant Rocks. *IEEE Pervasive Computing*, 2(4):2–4, October-December 2003.

Strohbach, M., Gellersen, H., Kortuem, G., and Kray, C. Intelligent Artefacts: An Embedded Systems Approach for Cooperative Assessment of Situations in the World. In *The Sixth International Conference on Ubiquitous Computing (Ubicomp 2004)*, Nottingham, England, September 7-10 2004.

Suchman, L.. Plans and situated actions. The problem of human-machine communication. Cambridge University Press, 1987.

Weiser. M.. The Computer for the 21st Century. *IEEE Pervasive Computing*, 1(1):18–25, January-March 2002. Reprinted from Scientific American, 1991.

Wiberg, M. and Ljungberg, F.. *Exploring the vision of "anytime, anywhere" in the context of mobile work*. Idea Group Publishing, 1999. Ed. Yogesh Malhotra

THE DIFFERENCE IS IN MESSAGING

Specifications, Properties and Gratifications Affecting the Japanese Wireless Service Evolution

Lars A. Knutsen[1] and Kalle Lyytinen[2]
[1]*Copenhagen Business School;* [2]*Weatherhead School of Management, CWRU*

Abstract: This paper extends common contentions of why the mobile Internet has been widely embraced in Japan but obtained lukewarm reaction in most GSM countries. In particular, we analyze commonalities and differences pertaining to the wireless killer application in both the West and the East – messaging. A framework consisting of service specifications, properties and gratifications is used to analyze short messaging, multimedia messaging and e-mail in Scandinavia and Japan. An architecture which better supports interlinking, integrating and transitioning of interpersonal and data-based communications over the service platform was successfully established in Japan while the disjointed nature of messaging, multimedia messaging and data services has inhibited Scandinavian users to fully embrace the mobile Internet. In Japan mobile e-mail integrated instrumental and aesthetic service properties on top of the powerful expressive service properties. Accordingly, content and process gratifications have augmented powerful social gratifications which initially have been driving m-service use in both places. Idiosyncrasies identified across service integration provide insights to critical enabling and constraining factors that shape development of mobile services.

Keywords: wireless services, mobile services, specifications, properties, gratifications, Scandinavia, Japan, messaging, SMS, MMS, e-mail, evolution

1. INTRODUCTION

Despite significant interest we have yet to answer what constitutes the main characteristics of a "killer application" in wireless services. Explanations with varying grounding abound why wireless service use blossom in Japan: the unique nature of the Japanese environment and the

architecture of NTT Docomo's i-mode service (Mitzukoshi et al. 2001; Sigurdson 2001; Baldi and Thaung 2002; MacDonald 2002; Ratliff 2002; Nielsen and Mahnke 2003; Elliott and Tang 2004; Sharma and Nakamura 2004). However, little attention is given to how messaging services – e.g. SMS (GSM) and mobile e-mail (Japan) - influence the growth of portal based data services. We lack knowledge of how service adoption success is related to particular specification features, service properties and gratifications. Past studies tend to categorize wireless services based on the content provided, i.e. news and entertainment (Baldi and Thaung 2002; Aarnio et al. 2002). While useful for statistical analyses such categorizing is problematic when wireless services are hybrids integrating services crossing multiple content categories (e.g. infotainment). The engorging variety of m-services and their interconnections are now increasingly blurring categorical service identification. Thus, we encounter trouble in identifying *the* killer application as it is part of a more holistic offering: "the *package* of content [emphasis added] that we put together was our killer application that helped i-mode take off" (former i-mode manager Toshiaharu Nishioka cited in Sharma and Nakamura 2004:47).

In this paper we propose that neither content variety nor service categorizing reveals the critical characteristics of a mobile 'killer application'. Rather, we move beyond categories to look deeper at specific services – those initially scaling and catalyzing mobile data use and their critical role in the service ecology. We focus on connections between messaging services and other types of data services and explain how the gestation of killer applications in Japan and Scandinavia differed. To this end we present a framework to analyze wireless services based upon service specifications, service properties and service gratifications.

2. GLOBAL KILLER APPLICATIONS: GROWTH OF SMS AND MOBILE E-MAIL.

The idea of a 'killer application' is widely disputed and a definition is needed (Lyman and Lowry 1996; Middleton 2003; Stroborn et al. 2004; Dey 2005). We refer to a killer application as *a distinct data service which has scaled, or is scaling towards dominance over other applications or services within a growing population of users and where its properties and name has achieved pervasive awareness.* As such, voice telephony falls outside our definition since it is not considered a data service. Connectivity is too vague to achieve pervasive awareness regarding the service name and properties. In contrast, news services or video and photo services fit with the definition. Though some of these services have reached pervasive awareness within

certain segments, their degree of meeting our criteria of pervasive awareness across a broader population is still nebulous. Only messaging services seem to fully comply with this definition.

2.1 Evolution of Mobile Data Services in Scandinavia and Japan

Scandinavia. Scandinavians are among the world's most eager users of SMS. Norwegians averaged almost 36 short messages per month during 2000 and have remained the most edacious 'texters' in Scandinavia until 2002. During 2002-2004 the Danes, probably due to intense price competition and introduction of flat rate pricing schemes, drastically leapfrogged their neighbors reaching an average of 102 short messages per subscriber per month during the first half of 2004. The telecommunications authorities of Sweden (PTS) reported that considerable higher prices on SMS have constrained SMS use to mirror that of its neighbors.

Table 1. SMS per subscriber per month in Norway, Denmark and Sweden

	2000	2001	2002	2003	1st half 2004
Norway	35.9	49.9	56.6	66.1	66.6
Denmark	21.5	30.7	39.9	71.6	102.1
Sweden	7.1	12.9	14.9	18.4	17.9

Research from Denmark (Constantiou et al. 2004) and Norway (Ling 2004) reveals that young users consume most data services. Ling (ibid.) reports of teens and adolescents sending between 6-9 messages per week, while people of ages 25-44 send between 1-3 messages per week. Yet, the text epidemic has not transmuted into much other data service use. Multimedia messaging (MMS) is starting to gain some momentum, but is still of miniscule significance.

Table 2. MMS per subscriber per month in Norway, Denmark and Sweden

	2002	2003	1st half 2004
Norway	0.03	0.39	1.00
Denmark	0.01	0.04	0.15
Sweden	0.02	0.06	0.19

WAP use remains absent and the growth of GPRS traffic is slow. Official GPRS traffic for 2004 show that the average Swedish and Danish GPRS subscriber generated respectively 0.64 and 0.45 MB in mobile data traffic over the first six months of 2004; equal to sending and receiving a few multimedia messages per subscriber.

Norwegian and Danish MNOs have recently experienced growth in premium SMS (PSMS) services offering voting and messaging in relation to

TV and radio shows, as well as purchasing of ringtones, logos, backgrounds and screensavers etc. No official statistics document the growth of content sales. However, a telecommunications analyst (2004:13) argues that the "PSMS market in Norway exploded" when a unified content provider access (CPA) platform was presented together with attractive revenue sharing schemes. In 2003 revenues from premium content was estimated in the range of 460-570 million NOK[1]. In Denmark, premium content service revenues were estimated to double from 180 million DKK in 2003 to 300-350 million in 2004[2]. PTS[3] claims that the basic service model for premium content services is also solid in Sweden, but that growth has yet to occur.

For all, however, the much expected growth in portal based m-services is still pending. Hence, new revenue sources have emerged incrementally around the SMS medium as extensions to social interaction based SMS use.

Japan. NTT DoCoMo and J-Phone were among the leaders in the amount of revenues derived from data traffic over cellular networks in 2002[4]. A considerable share of NTT DoCoMo's data revenues is attributable to content services. Although games, screen savers and ringing tones comprised more than 50% of i-mode user activity observers emphasize that "e-mail is the killer app" which "initially attracted users to i-mode" and which "remains critical to i-mode's appeal" (Mitzukoshi et al. 2001:93). For KDDI, the only operator in Japan utilizing WAP in its cdmaONE network, the data revenues of 2002 were tangent with the MNO average – approximately par with levels of Telenor of Norway.

Mobile e-mail is persistently identified as the most important data service (Hoffmann 2001) in Japan while SMS has small presence. Sigurdson (2001) reported that SMS - branded Cmail by KDDI – dominated the use of EZweb, KDDI's WAP based mobile Internet gateway, but has almost disappeared in Japan[5] with the dominance of e-mail.

Ishii (2004:48) suggests that "half of all mobile Internet users used email (excluding non-Internet short messages)" and that "the main usage of the mobile Internet is email" and Sharma and Nakamura (2004:49) argue that "Japan's growing band of e-mail junkies have driven the phenomenal growth in i-mode". On average, 87 e-mails were sent from a mobile phone on a monthly basis in Japan in 2002; 19 more than that what was sent over regular PCs (Ishii 2004). Although we were unable to find official statistics,

1 Økonomisk Rapport: http://www.orapp.no/oversikt/Argang_2003/8614/teknologi/8653

2 BT: http://www.bt.dk/mobil/artikel:aid=251258:fid=100300070/

3 PTS – Post og Telestyrelsen – report on Mobile Content: http://www.pts.se/Archive/Documents/SE/Slutrapport_mobilainnehallstjanster_2005_8_feb05.pdf

4 Nokia. 2003. A history of third generation networks: Nokia Networks. Available at: http://www.3gnewsroom.com/html/whitepapers/year_2003.shtml

5 NTT DoCoMo press release. NTT DoCoMo to Expand 3G International Short Messaging Service. Feb. 1st, 2005.

Mitzukoshi et al. (2001) reports of more than 100 messages sent and received per user per month in 2000. Given a 50/50 send-receive ratio, we find it plausible to expect similarities in the growth rates for mobile e-mail in Japan as witnessed with SMS in Norway and Denmark.

Younger users dominate data use (Funk 2001) with 50% of the users in their 20s or younger. Here 59% are males (59%) but 6% more females were present in the age group 19-24 (Ishii 2004).

Substantial growth in mobile Internet content, applications and multimedia communications is also evident in Japan (Funk 2001). In 2001, i-mode (NTT DoCoMo) had 30 million subscribers and J-sky (J-phone/Vodafone) and *au* (KDDI) each had 9 million. Currently, 86% of the near 85 million[6] Japanese mobile phone users subscribe to mobile Internet services. Next to e-mail the most accessed service types were: 1) search engines, 2) weather, 3) transportation information and maps, 4) music/concerts information, 5) news, 6) fortune telling, 7) sports, 8) computer games, 9) competitions prize/gift and 10) TV program/guide (Ishii 2004). Together data reported by Ishii (ibid.) Srivastava (2004) and NTT DoCoMo point to e-mail, video-games, ring-tones, information services and music as the most popular data services.

Mobile data revenues also boosted with the introduction of picture e-mail and mobile Java-applications. Picture mail was introduced by J-Phone in November 2000. In 2002, 60% of J-Phone subscribers owned a camera phone and subscribed to the sha-mail service which enabled sending and receiving of pictures and accessing of a 20MB personal photo gallery. These users generated on average 44% more revenue per user than a non-user (Nielsen and Mahnke 2003). A similar service was launched by KDDI in April 2002 and had 2.5 million subscribers by December 2002. NTT DoCoMo followed with the i-shot service in June 2002. By the end of the year 10% of its subscribers (4 mill.) had signed up for the service (ibid.). The different e-mail solutions have later been expanded to handle video-clips as well as other attachments (i.e. Microsoft Word and Adobe PDF).

In 2001, NTT DoCoMo introduced the i-appli service which enabled download and use of stand alone mobile Java applications (i.e. games, animations, karaoke, and tailored software applets). During 2001 the service gained 15 million subscribers. Witnessing that the average i-appli user generated double the data revenue of regular i-mode users (ibid.), KDDI and J-Phone introduced Java on their mobile Internet platforms in June 2001.

[6] Official statistics from the Telecommunications Authority of Japan: http://www.tca.or.jp/eng/database/daisu/index.html. Accessed March 2005.

2.2 Comparison of data service evolution in Scandinavia and Japan

While messaging is pivotal in Japan and Scandinavia, the access and use of data services differ. We recognize two paths in the use and innovation of mobile data services. First, after the WAP disappointment the Scandinavian approach concentrated on enhancing data services by innovating around SMS and MMS. Second, the Japanese operators have sought to innovate within and through the ecology of services which *integrate* e-mail as well as Internet based content delivery. Here, i-mode is most successful with offering access to 4,500 official and 84,000 unofficial m-service sites.

Although mobile e-mail is available in Scandinavia, it has been mainly offered as a client supplied by handset manufacturers and not as a operator-integrated solution. Statistics from Denmark[7] show that the SMS traffic at the only MNO offering integrated e-mail services, '3', is at par with the largest GSM operator. Hence, SMS' relative stronger technical and social pervasiveness (Knutsen & Overby, 2004) seem not only to hinder e-mail from substituting SMS but, as we now will analyze, may also hinder a similar development path as witnessed in Japan.

3. A FRAMEWORK FOR MOBILE SERVICE ANALYSIS

A tripartite approach is suggested to analyze the integrated nature of mobile services. First, service specifications as inscribed by the technical architects and corporate designers of those services/applications need to be identified. Second, we need to understand the expected and ascribed use properties of those services. Finally, we need to assess how the former two enable certain gratifications to be perceived by users.

Specifications refer to the specific texts characterizing the service as given by its architects – the objective inscriptions of technical details and capacities of a specific service. Examples of specifications[8] include payload, bit type characters (for alphabet), interface support, transfer and access support, format support etc. Their representations are manifests of the understandings and actions constituting the institutional environment putting them forth (Hargadon and Douglas 2001).

[7] Official Statistics from the Danish Ministry of Technology and Innovation.

[8] See for instance OMA specifications for MMS: http://www.wapforum.com/release_program/docs/mms/v1_2-20050301-a/oma-ereld-mms-v1_2-20050301-a.pdf

Properties convey the expected effects of service specifications upon humans or other artifacts. Three general categories which encompass the embodied symbol and virtual[9] service properties (Orlikowski 2000) that are subjectively constructed and re-constructed by actors interacting with the service or perceiving instantiations of it can be distinguished.

Instrumental properties encompass service capacity to support efficient interactions between humans and technologies. Examples include faster transmissions speed, augmented mass distribution capabilities, simpler text entry, reduced download/upload time, and reduced coordination time.

Aesthetic properties define the capability of a service to provide an aesthetically appealing human experience and/or enhance the appearance of other artifacts; e.g. evocation of visual, audible and physical sensing and decoration of artifacts that can convey the identity of the service or the identity of the service consumer.

Expressive properties signify the potential of a service to impress instantiations upon users and other artifacts; e.g. relative capacity of mobile service to pervasively establish connectivity and deliver content in a network of humans and artifacts.

Gratifications refer to pleasures, delights, and fulfillments users can perceive from using a mobile service based on their 'needs' and motivations (Blumler and Katz 1974; Cutler and Danowski 1980). Three main forms of gratifications can be outlined to understand how these effects can translate (or not translate) into perceived user value (Stafford et al. 2004).

Content gratifications encompass pleasures, delights and fulfillments experienced from the 'consumption' of messages offered through services; e.g. 'consumption' of content crafted by other users or service providers.

Process gratifications constitute a form of hedonic gratifications arising from the actual experience of using service (ibid.); i.e. playing, experimenting, and learning by exploring how services operate may itself be gratifying.

Social gratifications are gratifications perceived from service-enabled social interaction with physically or virtually co-present others; e.g. grooming of social relations, social coordination and organizing, social contact and identity expressing.

Specifications, properties and gratifications can be conceived to form the following relationships: 1) specifications have bearing upon properties and gratifications, whereas 2) properties influence the space of potential gratifications. Specifications are primarily constructed within the sphere of development, while properties and gratifications are culturally bound and

[9] For material artifacts, this would also include material properties (i.e. Orlikowski, 2000). Although dependent upon material artifacts, services are analytically distinguishable as virtual artifacts.

socially constructed valuations of technology primarily associated with the enactment of specifications. Hence, specifications mediate the creation of users' life-world experience with technology (Hill 1988). They become instantiated in the enactment enabled and constrained by modalities drawn upon to create meaning and legitimacy of technology use based on the monitoring of own and others direct and communicative actions pertaining to a particular technology (Blechar et al. 2005). Interpretations and enactments of properties and resulting gratifications enable isolated analysis of each tripartite element as well as assessment of their relationships to uncover what technical specifications and corresponding institutional arrangements enable and constrain interpretations of properties and the construction of gratifications. We can also analyze similarities and discrepancies in the 'text' of specifications and the degree to which the associated 'text' of properties and gratifications match and promote evolution of mobile service use.

4. ANALYZING MESSAGING TECHNOLOGIES USING THE TRIPARTITE MODEL

Specifications. As illustrated in table 3, SMS, MMS and mobile e-mail[10] differ widely their specifications off payload, text character limits, the formats supported, transmission protocols and addressing standards. SMS is by far the simplest service. It has a fraction of payload and character limit to that of MMS, only text format is supported and it depends on conversion for transfers to/from e-mail over Internet standards (e.g. HTTP). MMS has far richer specifications than SMS but is not as advanced as the mobile e-mail service. In particular, the payload of MMS (i.e. the 100kb limit for TDC customers in Denmark) is significantly less than that of mobile e-mail. In terms of formats supported, MMS and mobile e-mail are quite similar. Both support several multimedia formats and HTML. However, the attachment function of mobile e-mail opens for transmission of other file formats. With respect to addressing SMS and MMS support the E-164 phone number standards. MMS also supports the RFC 2822 standard e-mail addressing. Mobile e-mail, on the other hand, primarily supports the latter and includes, if not changed by the user, the phone number as part of the default e-mail address. Finally, the transmission protocols for the respective services differ considerably. While MMS utilizes either WAP WSP or HTTP/TCP/IP, mobile e-mail utilizes the latter and SMS utilizes X.25 which demands a

[10] We draw upon the mobile e-mail services offered by NTT DoCoMo over 2G and 3G networks as these have the strongest manifest in Japan.

message router appliance in order to convert and transfer data over the other transfer protocols. Importantly, MMS uses a different access point name (APN) than WAP and this disables the simultaneous use of WAP and MMS.

Table 3. Comparison of selected specifications of SMS, MMS and Mobile e-mail

	SMS	MMS	Mobile e-mail
Maximum payload	140 bytes	100 kb	500 kb
Text character limit	160 (7-bit), 140 (8-bit), 70 (double-byte)	Limited to maximum payload enabled by operator, MMSC and handset.	i-mode – i-mode: 4,000/2,000 Foma – Foma: 10,000/5,000
Formats supported	Text	Text, MPEG4 HTML Gif / Animated GIF JPEG, WBMP, PNG, MP3	Text, MPEG4, HTML, GIF, JPEG, WBMP, PNG, MSWord, Excel, PowerPoint, Adobe PDF
Transmission protocols	X.25	WAP WSP stack or HTTP/TCP/IP WAP push	HTTP/TCP/IP
Addressing standards	164 phone numbers	E.164 phone numbers and RFC 2822 e-mail addresses	RFC 2822 e-mail addresses

Properties. The specifications yield variations in types and degrees of instrumental, aesthetic and expressive service properties (table 3).

Instrumental properties. The capacity of the messaging services to promote efficient interactions between human beings and other artifacts are similar but also idiosyncratic. First, one-to-one and one-to-many properties to quickly exchange messages with other networked people are shared. But a central difference is that MMS and mobile e-mail better facilitate ease in horizontal network bridging between mobile phone users and computer based e-mail users. With respect to efficiently facilitate interactions we therefore see augmented properties for MMS and mobile e-mail to assist social micro-coordination (Ling 2004).

Second, mobile e-mail and MMS have capacity to supply content to other users through the support of HTML and multimedia. Initially limited to text and images (incl. animated), MMS now also supports video-clips, and the MP3 audio format (e.g. for sharing of recordings and podcasting). However, unable to support other third party file formats leaves an advantage to e-mail in the ability to instrumentally support a wider range of everyday activities.

Third, all three support instrumental interactions between artifacts: configuration (e.g. installation/set-up scripts) as well as service specific scripts (e.g. for payment, ticketing, and content purchases). However, the

richer format support of MMS and mobile e-mail offer wider opportunities to support exchanges between commerce services and other applications.

Aesthetic properties. Much of the success of content personalization, i.e. PSMS content in Norway and ringtone purchases over i-mode in Japan, are due to augmented aesthetic properties. Personalization content supplied over SMS can potentially enhance the aesthetic effect of messages. Generation txt's invention of novel use of alphanumeric characters (e.g., smileys :-) to display emotions (Rheingold 2002; Ling 2004) is one example of aesthetic enrichment of SMS. However, the specifications of SMS severely constrain the continuation of such aesthetically oriented innovations – e.g. towards animated GIF emoticons. Thus, extended multimedia capabilities and HTML add dimensions to the virtual aesthetics beyond that of SMS by enhancing content tailoring. For instance, NTT DoCoMo increased mobile e-mail attachment payloads at the same time it enhanced the Deco-mail message tailoring specifications ; an example of how enriched aesthetic service properties can enhance aestheticism of communicative interaction content.

Expressive properties. The expressive properties are high for all three as they powerfully, pervasively and rapidly can diffuse messages in enormous user networks. The expressive latitude to construct messages may however be lesser for SMS due to the limited number of character support. Although Ling (2004) found that character limitations of SMS are seldom reached in practice, the vast use of abbreviated language (Rheingold 2002) may also suggest that there are situations in which the increased payload capacity of MMS and e-mail can offer more advantageous expressive properties.

The messaging technologies exhibit network effects (Katz and Shapiro 1985; Katz and Shapiro 1986) since each becomes more valuable as the number of users increases. While content supply tend to scale according to Sarnoff's law, messaging services have properties to scale exponentially according to Metcalfe's law for two-way communications services and Reed's law for group forming networks (Reed 2001). Here, SMS and e-mail has the advantage over MMS due to the enormity of current users. With approximately two thirds of the world's wireless users being GSM customers and the fact that SMS is also available over other networks[11] – even the Japanese – it may even hold stronger expressive properties than that of mobile e-mail in terms of pervasive message delivery. Yet, if we consider the enormous network of computer based e-mail users around the world[12], another powerful network externality enabled by mobile e-mail with other TCP/IP based networks emerges giving an upper hand to mobile e-mail.

[11] The GSM Association, www.gsmworld.com

[12] ITU – www.itu.org - reported of close to 700 million Internet users (in terms of subscribers) in 2003. It is likely that the number of e-mail users is far larger than this.

In content delivery, the second element of expressive properties, mobile e-mail is stronger. This results from enhanced multimedia capabilities and the capacity of HTML to link to other types of content. NTT DoCoMo incorporated 'send to' and 'web to' capabilities with e-mail enabling users to hyperlink to content sites in message exchanges. This enable a helix effect, a form of positive feedback effect (Funk 2001; Lee and O'Connor 2003), which augments the expressive properties of mobile e-mail. Technically MMS has similar capabilities. But relatively limited MMS use and lack of mobile content sites constrains the coiling of such effects.

4.1 Gratifications

Content gratifications. The variety of content subject for messaging offers a tremendously versatile set of possible pleasures, delights and fulfillments through consumption. Content used to support individualized interaction and social relations has been found to be the widely used and are probably the most gratifying. Of the 882 SMS messages investigated in Norway, Ling (2004) found that about two thirds contained simple statements most often associated with micro/social-coordination (33%) followed by grooming messages (17%) – nurturing of friendships, relations and romances – answers (14%), questions (11%), information (6%), and personal news (5%). Analogous, a majority of e-mails sent from mobile phones in Japan were also directed to a limited number of people in immediate vicinity and with whom face-to-face interactions are frequent (Ishii 2004)[13]. A former i-mode manager, Toshiharu Nishioka, has also stated: "Sending short messages, such as saying goodnight to a friend, is one of the most popular uses of i-mode" (Sharma and Nakamura 2004)[14]. This suggests the personal specificity and adaptability of messaging content is superior to what any third-party information, entertainment or location sensitive service provider can offer. However, as witnessed with sha-mail (Ishii 2004), third party content providers can enhance such gratifications further by offering graphical content that increase use of endearing, emotional and experiential content; e.g., emoticons, pictures and sound. If MMS prices approximate those of SMS we may thus see a similar path in Scandinavia. As of now, however, SMS does not appear to be severely disadvantaged with respect to the prevailing type of gratifying contents.

Process gratifications. There are two contrasting issues pertaining to process gratifications. On one hand, efficiency and simplicity can be the most gratifying process aspect. However, such instrumentally oriented fulfillments can sometimes be foreshadowed by gratifications derived from

[13] Although there is an indication concerning this, this issue warrants further studies.

[14] Notice that messaging is here considered an inherent and inseparable function of i-mode.

carefully crafting an aesthetically appealing message. Ling's (2004) research provides empirical support: younger users, men in particular, tend to write short messages. The average was 5.54 words for men and 6.32 for women. Women also appear to take more care in crafting messages with appropriate punctuation and capitalization and emoticons were more commonly used by women of ages 13-25. Not only does this signify that women's relatively stronger social interaction skills can be instantiated in messaging (ibid.), but also that women in general receive higher gratifications from advanced messaging. Interestingly, the messaging technologies (MMS and e-mail more than SMS) can support both ends of the simple-complex spectrum. Nevertheless, since the enhanced multimedia capabilities of MMS and mobile e-mail do not compromise process efficiency, their enriched multimedia capabilities can also offer complementary process gratifications for both genders. Importantly, process gratifications gained from crafting own content appear to be larger than process gratifications gained from 'surfing', 'browsing', or 'clicking' the mobile Internet.

Social gratifications. Messaging weaves new socio-communicative textures of contextual interaction. Although taking place culturally distinct ways, several commonalities in social gratifications can be identified. In both Japan and Norway messaging is involved in establishing and maintaining relationships. Ling (2004) reports of grooming as well as terminations of relationships via SMS in Norway and Ishii's (2004) research in Japan document that mobile e-mail provides a double fulfillment in both supporting communication of personal feelings while also securing that face-to-face interaction can be minimized. Messaging also fulfills an important role in breaking down barriers between people in the beginning of relationships, help in the transitioning to synchronous communication as well as erect barriers if one party chooses to disengage (Ling, 2004). Although most social gratifications arise from everyday person-to-person communications, i.e. (boy/girl-) friends, spouse, other family members, and peers, the versatility of mobile messaging also introduces hybrid forms of inter-personal communications. This is significant for social gratifications related to the nurturing of group relations as well as for understanding how group forming networks operate. Not only can they scale exponentially (Reed 2001), but they can have a potent effect on the velocity at which messages traverse and how messages bind or unbind people's social ties. Research shows that crafting as well as consuming of messages can take place as a social activity among co-present people, i.e. local social interaction in sharing and reading aloud, passing digital devices around etc. (Ling 2004). Hence, gratifications associated with group membership, feelings of belonging, participation and identity across wider time and space spans are enabled by mobile messaging. Research in Denmark supports this

in finding that independence of time and space (21%), contact with friends and peers (21%) and contact with family members (17%) are top three contributions of mobile service activities (Constantiou et al. 2004). While cultural idiosyncrasies exist, e.g. limited private space as a conduit for messaging success in Japan (Sharma and Nakamura 2004), the freedom of being able to connect socially while simultaneously disconnect from constraints of physical space seem fundamental to social gratifications sought in both geographical areas.

SMS, MMS and mobile e-mail all strongly enable social gratifications. Few details exist if MMS brings social gratifications beyond that of SMS, but the rapid adoption and use of animation, photo and video documented with mobile e-mail in Japan (Nielsen and Mahnke 2003; Ishii 2004) suggest that sending and receiving of aesthetically enhanced content augments social gratifications by offering new content for communication; e.g. interchange of personal recordings and pictures, and moblogging where users post pictures, messages and recordings from mobile phones onto personal Internet pages (Brown 2004). Also, the ability to use 'mail to' and 'web to' type functions and link to content enable new integrations can expand interpersonal communication and the magnitude of social gratifications.

5. E-MAIL: AN ARBITER OF DATA SERVICE USE

The differences in specifications in messaging yield different scope in properties and gratifications enabled. Despite striking similarities in properties and gratifications associated with text based messaging, Japan is experiencing greater use of data services and thus greater cultural and social embodiment of additional gratifications. So why does multimedia messaging and mobile Internet use still differ? Our analysis points to e-mail as a killer application enabling evolution in a larger socio-technical configuration.

First, NTT DoCoMo created a integrated solution which did not erect barriers between messaging and content browsing by using de-facto Internet standards (Funk 2001). This promoted seamless service migration and unleashed the power of content and hyperlink-interchange in messaging. A natural connection between data services and mobile e-mail was established as an imperative gene promoting service-synthesis. Mobile e-mail became the killer application gluing together a service-smorgasbord that embodied a broad set of instrumental, aesthetic and expressive properties. Hence, e-mail yielded gratifications *as a whole* beyond any other free-standing data service (e.g news, entertainment etc.) due to its superior aesthetic and expressive service properties and strong baseline ability to provide social gratifications.

Second, the use of e-mail together with packet based technology from the inception of i-mode offered advantages in perpetual connectivity and scalability. A more flexible pricing scheme calculated based upon the size of messages made users not choose either between SMS or MMS price, but based upon the relative gratification perceived during composition and the imagined gratifications of the receiving person(s). In contrast, MMS is offered as a separate service and is charged per message. This forces new acronyms, learning and choices regarding how a service relates to the scripts and meanings currently known instead of evoking "interpretations among potential adopters that are based on adopters' past understandings and experiences" (Hargadon and Douglas 2001:478).

Third, while consumer oriented cultural explanations (e.g. Baldi and Thaung 2002) are "partly true – but mostly false" (2004:46) institutional and cultural differences in business organization related to wireless services are important. One of the master-minds behind i-mode, Takeshi Natsuno, proclaimed that: "The true mechanism of the great success is that the operator has made a function of coordination of the total value chain" (Sharma and Nakamura 2004). Such organizing sharply contrasts with the more "silver bullet" based vertical orientation adopted by major players in the West. Our analysis suggests that the former approach appears to facilitate horizontal and vertical bridging of specifications and properties as well as an extended line of achievable gratifications.

6. CONCLUSION

The rift erected between SMS and MMS breaks down a natural path to extended service use, while the original Japanese architecture promotes an incremental evolution which promote advanced service use through a robust architectural design which is flexible and rigid as well as extensible and simple (Hargadon and Douglas 2001). Smoreda and Thomas (2001) suggested that the Internet and SMS should be brought together in order to "further stimulate the use of SMS". In applying the framework above we find that e-mail forms the most promising technology to facilitate this bridging. Other messaging services have weaker properties to function as the socio-technical glue binding data services together. The GSM Association recently initiated the "Integrated Messaging Initiative"[15] aiming to create an integrated user experience for SMS, MMS and e-mail. If our analyses hold water, this initiative should be focused on migration towards e-mail.

[15] GSM Association. February 15th, 2005. Press release: Integrated Messaging Initiative Will Drive Growth of Richer Messaging Services. www.gsmworld.com

However, in being recalcitrant to rock the profits of the SMS juggernaut by following the empirically documented desire for e-mail (Anckar and D'Incau 2002; Constantiou et al. 2004) operators are now inviting new players to the scene. Three million Blackberries with primary functions being e-mail and calendar have been sold in the US in 2004 – a traditionally slow messaging market – and Blackberries are increasingly gaining patronization in Europe (Gibbs 2005). Revenue hungry operators have the opportunity unleash the power of mobile e-mail to grow data service use. But this requires operators, handset providers and content providers to commit to interorganizational revenue sharing and innovation (Kodama 1999; Funk 2001; Mitzukoshi et al. 2001; Baldi and Thaung 2002; Ratliff 2002) so that value creation both at the supply and demand sides become sufficiently lucid and geared towards economies and gratifications of scale.

ACKNOWLEDGEMENTS

This paper partly results from research conducted in the Mobiconomy project at Copenhagen Business School. Mobiconomy is partially supported by the Danish Research Agency, grant number 2054-03-0004.

References

Anckar, B. and D. D'Incau (2002). Value creation in mobile commerce: Findings from a consumer survey. *JITTA : Journal of IT Theory and Application* **4**(1): 43-65.

Baldi, S. and H. P.-P. Thaung (2002). The Entertaining Way to M-Commerce: Japan's Approach to the Mobile Internet - A Model for Europe. *Electronic Markets* **12**(1): 6-13.

Blechar, J., et al. (2005). Reflexivity, the social actor and m-service domestication: Linking the human, technological and contextual. *IFIP 8.2 WC,* Cleveland, OH, USA.

Blumler, J. G. and E. Katz, Eds. (1974). *The uses of mass communications: Current perspectives on gratifications research.* Beverly Hills, CA, Sage.

Brown, K. (2004). Making Money From Moblogging. *Wireless Week,* Reed Business Information. **10:** 16.

Constantiou, I. D., et al. (2004). Strategic planning for mobile services adoption and diffusion: Empirical evidence from the Danish market. *MOBIS 2004 - IFIP TC8 Working Conference,* Oslo, Norway, Kluwer.

Cutler, N. E. and J. A. Danowski (1980). Process gratifications in aging cohorts. *Journalism Quarterly* **57**(Summer): 269-277.

Dey, S. (2005). The Evolution of 3G Wireless Data Services. *Byte.com,* CMP Media LLC: N.PAG.

Elliott, G. and H. Tang (2004). The wireless mobile internet: an international and historical comparison of the European and American wireless application protocol (WAP) and the Japanese iMode service. *International Journal of IT & Management.* **3:** 1.

Funk, J. (2001). *The Mobile Internet: How Japan Dialled Up and the West Disconnected.* Kent, UK, ISI Publications.

Gibbs, C. (2005). E-mail companies target casual user. *RCR Wireless News*, Crain Communications Inc. (MI). **24:** 6.

Hammond, K. (2001). B2C e-Commerce 2000-2010: What Experts Predict. *Business Strategy Review* **12**(1): 43-50.

Hargadon, A. and Y. Douglas (2001). When Innovations Meet Institutions: Edison and the Design of the Electric Light. *Administrative Science Quarterly* **46**(3): 476-501.

Hill, S. (1988). *The Tragedy of Technology*. London, Pluto Press.

Hoffmann, A. (2001). The Other i-modes. Fifteen million happy non-DoCoMo users can't all be wrong. *J@pan Inc.* **20:** 60.

Ishii, K. (2004).Internet use via mobile phone in Japan.*Telecommunications Policy*.**28:** 43-58.

Katz, M. L. and C. Shapiro (1985). Network Externalities, Competition, and Compatibility. American Economic Review **75**(3): 424-440.

Katz, M. L. and C. Shapiro (1986). Technology Adoption in the Presence of Network Externalities. *Journal of Political Economy* **94**(4): 822-841.

Kodama, M. (1999). Business innovation through joint ventures supported by major businesses. *Journal of Management Development* **18**(7/8): 614.

Lee, Y. and G. C. O'Connor (2003). New Product Launch Strategy for Network Effects Products. *Journal of the Academy of Marketing Science* **31**(3): 241-255.

Ling, R. (2004). *The Mobile Connection. The Cell Phone's Impact on Society.*, Morgan Kaufmann Publishers.

Lyman, P. and C. B. Lowry (1996). Access is the killer application. *Journal of Academic Librarianship*, Elsevier Science Publishing Company, Inc. **22:** 371-375.

MacDonald, D. J. (2002). NTT DoCoMO's i-mode: Developing Win-Win Relationships for Mobile Commerce. *Mobile Commerce. Technology, Theory and Applications.* B. E. Mennecke and T. J. Strader. London, UK, Idea Group Publishing. **1:** 1-25.

Middleton, C. A. (2003). What if there is no killer application? An exploration of a user-centric perspective on broadband. *Journal of IT,* Routledge, Ltd. **18:** 231-245.

Mitzukoshi, Y., et al. (2001). Lessons from Japan. *Telephony* **240**(3): 92-95.

Nielsen, L. E. and V. Mahnke (2003). The challenging of the old Tsunami: the case of NTT DoCoMo: Competitive, regulatory, innovation, network and overseas challenges -. Working paper, nr.2003-06. Copenhagen, Copenhagen Business School.

Orlikowski, W. J. (2000). Using technology and constituting structures: A practice lens for studying technology in organizations. *Organization Science* **11**(4): 404-428.

Ratliff, J. M. (2002). NTT DoCoMo and Its i-mode Success: origins and implications. *California Management Review,* California Management Review. **44:** 55-70.

Reed, D. P. (2001). The Law of the Pack. Harvard Business Review **79**(2): 23-25.

Rheingold, H. (2002). *Smart Mobs. The Next Social Revolution. Transforming Cultures and Communities in the Age of Instant Access.* Cambridge, MA, Basic Books.

Sharma, C. and Y. Nakamura (2004). The DoCoMo Mojo. *J@pan Inc.*, LINC Media, Inc. **3:** 44-49.

Sigurdson, J. (2001). WAP OFF - Origin, Failure and Future. Unpublished Working Paper.

Srivastava, L. (2004). Japan's ubiquitous mobile information society. *info* **14**(4): 234-251.

Stafford, T. F., et al. (2004). Determining Uses and Gratifications for the Internet. *Decision Sciences* **35**(2): 259-287.

Strand, J. (2004). Norwegian operators push premium SMS. *New Media Age*, Centaur Communications: 13.

Stroborn, K., et al. (2004). Internet payments in Germany: a classificatory framework and empirical evidence. *Journal of Business Research*, Elsevier, Inc. **57:** 1431-1437.

Aarnio, A., et al. (2002). *Adoption and Use of Mobile Services. Empirical Evidence from a Finnish Survey.* (HICSS-35'02), Big Island, Hawaii.

UNDERSTANDING THE USER WITHIN THE INNOVATION SPIRAL

Reginald Coutts[1], Pamela Coutts[2] and Kate Alport[1]
[1]Coutts Communications, 30-32 Stirling St, Thebarton, SA 5031 and University of Adelaide, Nth Terrace Adelaide SA;[2]Coutts Communications;

Abstract: This paper will examine the concept of an innovation spiral process in relationship to the mobile communications sector of the Information and Communications Technology (ICT) industry which is the product of convergence of the telecommunications and the information technology (IT) industries. The fundamental theoretical framework for the paper is that innovative applications prescribed by users of an adopted technology can be a significant driver of further product evolution which then can fuel further market innovation. New products that are spectacularly 'successful' are those that give rise to this spiral of product innovation and market innovation. The paper will first review the macro perspective of this theoretical framework by analysing the spectacular success of the short message service (SMS) in the GSM digital mobile communications over the last twenty years from the product concept. This conceptual framework is then considered at a micro level to review three consecutive research projects by the authors over the last ten years. The broad aim of these projects was to better understand potential user adoption of new mobile telecommunications products. The first research project in the mid 1990's examined the barriers and enablers to the adoption of mobile phones by selected disadvantaged groups in society, in particular, people with disabilities. A modified focus group methodology based on interactive workshops was developed from this research project to gain insights into user innovation. This methodology was developed further in more recent research projects looking at the likely user take up of evolving multimedia capable mobile devices for innovative applications. The second study indicated that the market evolution of mobile internet like applications is likely to be very different from those developed for the fixed internet because of the different characteristics of the user groups. While the individual research projects have been published, this paper brings together the macro and micro perspectives of this innovation spiral to demonstrate the value of this theoretical framework for market forecasting for realising technology commercialisation. The research also has implications for the 'form' of new

high technology products and how they are marketed which places less emphasis on technical features but more on matching with user needs.

Keywords: innovation; telecommunications; mobile communications; GSM; SMS; qualitative research; multimedia; commercialization

1. INTRODUCTION

There has been much written about 'disruptive technologies' (Christensen, 1996) in the last nine years particularly with reference to the dramatic effects of innovations in ICT technologies which have reshaped economies, companies and the way people live. It is clear that new perspectives on how innovation evolves are still required to give deeper insight into how to better-forecast future technology adoption scenarios.

Bar has described the evolutionary nature of the innovation process that "*highlights the existence of feedback loops, interactions and linkages between users and producers*" (Bar and Riis 2000). For complex network based service products like in the telecommunications sector, Damsgaard develops a new framework combining technology evolution and service adoption that recognises the distinct but interactive role of infrastructure and service innovation described as "*A self-enforcing spiral of mutual re-enforcement of both infrastructure innovation and innovation adoption can unfold under favourable circumstances*" (Damsgaard and Gao 2004). This paper extends this framework to consider distinct phases of innovation.

While this paper does not pretend to offer the answers to the forecasting of the market or user demand for ICT technologies, it does offer a new perspective on how innovation gathers pace in distinct phases and the key role of users in this process. The process in each of these phases has a different relative dynamic of market and product innovation which we call the innovation spiral.

To illustrate the conceptual utility of this evolution framework, the paper considers the history of the evolution of the SMS service in GSM that is generally regarded as one of the most successful innovations in the mobile sector of telecommunications in the last ten years. The evolution of the product and the market are seen as occurring in four phases. The 'innovation spiral' as we have termed to describe the dynamic has distinct characteristics in each different phase. This could be considered the 'macro' view of this innovation evolution process.

The paper then examines three successive user focused research projects by the authors (Coutts, 1998, 2002, 2003) conducted over the last seven years in the context of this conceptual framework for understanding the lay

users role in mobile technology evolution. It has been noted that lay users "*often think outside of the box and imagine potential applications that are not simply extensions of existing practices*" (Bar and Riis 2000). Evidence suggests that the "*tight relationships between users and producers from early stages of research are especially critical where radical innovation is concerned*" (Bar and Riis 2000). The methodology developed during this research in a sense provides a window into how users appropriate and drive innovation. Experience with a modified focus group methodology that significantly improves the ability to develop insights into how the user innovates, can indicate the likely developmental phases of a technology. We define this process as user innovation in that it inspires innovation by the industry to modify the product or create a new one!

The last part of this paper brings the two macro and micro perspectives together to illustrate the utility of the innovation spiral and a methodology in developing more informed scenarios of future demand for ICT product innovations. The overall recommendation arising from this research and the proposed framework is for operators to undertake regular engagement with the lay user community in order to monitor and in a sense calibrate the innovation evolution of the range of products.

2. TECHNOLOGY INNOVATION FRAMEWORK

The conceptual framework to be developed in this paper is that *successful* technology innovation can be viewed as going through distinct phases where at one or more phases the 'innovation spirals' when the market rapidly expands in an 'unpredictable way'. This framework builds on the concept of a "*self-enforcing spiral*" coined by Damsgaard (Damsgaard and Gao 2004) but with the aim of gaining insights from lay users about how the technology may evolve.

In the initial phase, at the onset of innovation, there is both market and product uncertainty (Utterback et al 1995) and is the point before any innovation spiral, as we define it can begin to occur or *turn*. The phases of interactive innovation start when the service is available to users.

For the innovation spiral to occur, users or as it is otherwise termed, a *market segment* of a technology redefine the utility of the technology through experience, giving rise to further technology improvement or even giving rise to a new variant of the technology. This user inspired innovation fuels even greater user adoption. The innovation spiralling effect can accelerate in distinct phases in the product lifecycle that results in what 'appears' to be a discontinuity. In reality the change is continuous, even

though adoption occurs at a different rate. The challenge of the developed modified focus group approach discussed in section 4 is to gain key insights into these potential inflexion points of increased innovation and market adoption.

Firstly, it is necessary to define our terms. Technology can be understood as "*the knowledge embodied in human action to achieve practical results*". *Invention* is the creation of a new idea or concept whereas *innovation* is *the process of turning the new concept into commercial success or widespread use*. This paper is primarily concerned with technological innovations which are those with a significant "performance content". However, such 'performance content' must be able to be realised in increased market value.

The proposed hypothesis is that users of technology can reinterpret the utility of a technology considered in a way that the industry does not anticipate and that an understanding of that process can inform market evolution forecasting. This 'innovation spiral' has been examined in the context of technologies in the mobile sector of the telecommunications industry with reference to the history of the short messaging service (SMS) on mobile phones. In the case of the history of SMS, the market or user experience of the early form of SMS spurred a response, albeit rather belatedly, by the industry to explore and develop the technological and commercial means of providing an improved capability. It is argued that n this way the users are functioning as innovators which then promotes an innovative technological response from the industry. In this spiralling pattern of innovation the user is doing much more than just being stimulated to buy and use a new service, they are actually adapting the existing technology and effectively creating a new technology . In a sense this series of innovation spirals and can be considered as distinct re-tuning of the familiar market adoption S curve as the product better meets market demand.

This process of users reinterpreting or creating a new context for a communications technology platform was first observed by the authors (Coutts 1998) in their research with workers in the community service sector supporting people with disabilities. Despite the assumed high utility of a mobile service for a workforce that was undergoing decentralisation and increasingly working out of the office under time pressure, take up of mobiles was very low. In their work with this sector the researchers were in fact promoting and observing the process of user innovation – a process that usually takes place unobserved and unscripted, unchoreographed or recorded, let alone interpreted, in a way that would ultimately make it technically and commercially viable. This should be distinguished from the approach known as 'activity testing' which is the observation of users for the purpose of fine-tuning an existing technology capability in the context of intended use by the technology supply industry.

The focus of this paper will be based on the user inspired innovation within the same market, in this case Australia. However, the supply side of the industry also is inspired by innovations in other outside markets where it is *perceived* there are sufficient similarities.

3. SMS – A MACRO CASE STUDY

The Short Message Service (SMS) is an integral part of the digital mobile technology GSM that was launched in 1992 in Europe but now extends over most of the world. Prior to the GSM based mobile service, most markets had experience with an analogue mobile service that just offered a mobile voice telephony service. While SMS was originally only a minor 'value added feature' of the GSM mobile voice service, similar to the then paging service, it has evolved over the last 10 to 15 years to become a mass communications service in its own right.

In the last 5 years SMS and *premium SMS* as it is called supports a whole diversity of business applications ranging from managing field staff to enabling television viewers to vote in 'reality TV.' There is no question that SMS, coined the "ugly duckling" of GSM (Trosby 2004) has been an outstanding market success that could not have been predicted. Many business strategy analysts would argue SMS is an example of a "discontinuous" or "disruptive technology" (Christensen 1996) that has changed the communications paradigm. The perspective in this paper is that SMS is a 'continuous' technology but can be viewed as undergoing changes in quite distinctive phases each identified as ***a spiral of innovation*** along a continuum but with different characteristics.

The first stage of SMS evolution we would term **Phase 0** as this was the pre-launch stage. This included: the specification of the three service elements in Europe in 1987, how these requirements would be incorporated in the GSM architecture through and the launch of commercial services in Australia in 1993 one year after Europe. Some key decisions were taken during this phase by visionary engineers with experience of data communications in the fixed telecommunications network. The expectation by most outside this inner group was that FAX and circuit switched data were the significant non-voice services to be concerned about. Phase 0 corresponds to the 'onset' of innovation (Utterback and Afuah 1995) where there is great uncertainty about the product and the market.

Phase 1 of the SMS service or the first turn of the spiral of innovation in Australia began in 1992 with the launch of the first commercial service. This phase can be categorised as the 'technology innovation only' phase. SMS was positioned as similar to the *existing* product, the paging service, but

being differentiated by its integration with the mobile phone[1]. SMS was an alternative to diversion to voice mail[2] for digital mobile phones. Mobile phones initially in this phase did not support two way SMS. While international roaming was a key feature of the GSM voice service, SMS international roaming was sporadic due to the incompatibility of the SMS Service Centres. Thus SMS technology in this phase was well developed and stable but not marketed and not generally adopted in the market. By the end of this phase in 1996, two way SMS and international SMS to support international roaming were in place, but SMS messaging between users using mobiles on different networks was not supported by the carriers. SMS messages could however be sent between networks via the internet developed by innovation third players but could not be charged for! This restriction from a user's perspective was a significant usage barrier to SMS being used as a distinct messaging platform rather than just an adjunct to the voice service.

Phase 2 of the SMS evolution is where SMS exploded as a service in its own right. This we would argue was because of strong feedback in the innovation spiral from users feeding ongoing technology innovation. This explosion was ignited first in 1996 by the introduction of the pre-paid mobile service in Italy which in turn influenced the operators in Europe and Australia. Pre-paid technology removed the significant barrier to the development of the youth market that had been deterred from entering the market, by credit controls and budget concerns. However, the current billing system could not charge users for SMS and inter network SMS. Thus phase 2 was a key *process innovation* enabling growth in what, till then, was a SMS service *product innovation* as described in the innovation literature.

Thus this phase of intense user and technology innovation saw several key market and technology driver elements converge:

- Mobile youth hungry to communicate cost effectively became the new market driver with pre-paid mobile phones
- The initial billing system's inability to charge for many SMS messages from pre-paid mobile phones or where inter network SMS meant that SMS at zero charge was VERY cost effective for this new youth market
- The introduction of 'predictive text' messaging which increased the SMS usage per user significantly[3]

This was a boom period for SMS was characterised by *both* 'user innovation' and 'technology innovation' as operators quickly implemented effective charging for SMS messaging from both pre-paid and inter-network

[1] In the US it was common for users to carry a pager as well as a mobile phone because of the 'mobile party pays' charging principle which was unlike most of the world.

[2] Diversion to voice mail was very common for the previous analogue mobile service.

[3] One estimate is the 'predictive text' increased usage by 30% per user

SMS mobiles. In Figure 3.1 shows the dramatic increase in SMS traffic when it was marketed to the youth market at zero charge for pre-paid mobile originated SMS. When charging was implemented, usage dropped but when inter-network SMS was activated with charging, the SMS traffic started to climb dramatically. The 'network externality value' to users of being able to SMS anyone irrespective of the network provider far exceeded the SMS charge of 20c per message.

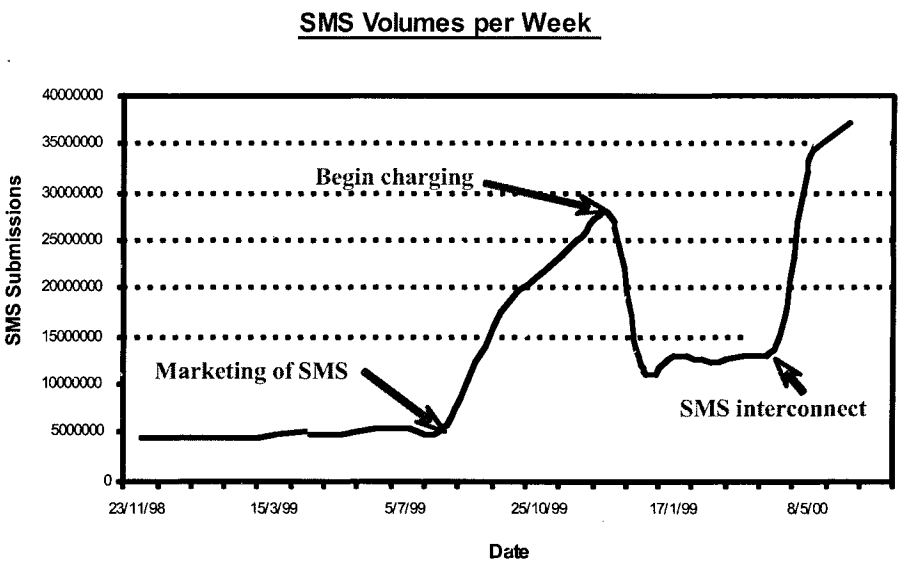


Figure 1. Impact of Inter-network Connection on SMS Traffic in Australia

During the rest of this phase operators did not lower their charges for SMS as they were having extreme difficulty coping with the traffic volumes. The figure shows that the volume increased by an order of magnitude in some 18 months with greatly increased usage per user. This growth in SMS in Australia is shown in Figure 3.2.

By the end of this phase SMS had become a phenomenon particularly associated with youth but its appeal across age demographics had begun and operators were beginning to have to reduce SMS charging as it became a commodity. Since SMS messaging was transparent to the users network operator, operators began competing on the price of SMS messaging. Phase 3 was primarily user inspired innovation by youth consumer users

appropriating SMS for their use so SMS became a new cost effective communications cultural phenomenon. Research in Japan illustrates the insights from an anthropological perspective of the market adoption of messaging services (Ito 2003).

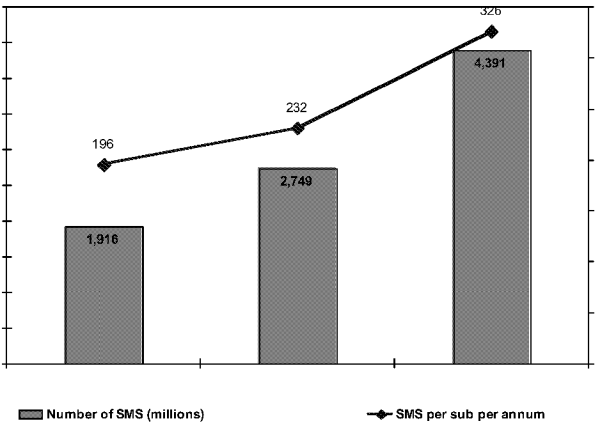


Figure 2. Growth of SMS in Phase 2 in Australia

As this phase in SMS development paralleled the spectacular growth of the internet, attempts by the industry to foist the promise of a "mobile internet" on users with the introduction of Wireless Application Protocol (WAP) was generally regarded as a failure. Another example naively compared to SMS is MMS (Multimedia Service). While cameras integrated into mobile phones has been extremely popular, MMS has been very poorly received generally by the market as it is *very* different to SMS from a user perspective and has been poorly marketed. Developing a new value proposition linking the power of the internet in the user context of mobility requires a user centred framework focused on developing successive valued applications in evolved phases. We would argue i-Mode developed by NTT

DoCoMo in Japan in this same period was a success *because* it focused on market innovation around applications rather than on technology innovation. The i-Mode model presents a total value proposition to the market and not a technology features jig saw puzzle. Important in the overall i-Mode business model was a viable business model for content developers which was not the case for WAP model in most countries

Phase 3 for the SMS innovation spiral started in Australia in about 2001 the overall technology had matured so that it included the useability of SMS on phones as well as the network technologies of billing and robust service centres. SMS pricing from operators was falling enabling SMS to become a platform for messaging applications by business. This phase is characterised by a combination of technology innovation and user innovation but within the business sector. Users in this context are intermediate users or a business. With reduced pricing, particularly for wholesale purchase as a result of the commoditisation of SMS in the previous phase, market innovation in the financial, retail and business service sectors started to flourish and still is expanding. The introduction of premium SMS at the end of Phase 2 with abbreviated numbers enabled the growth of interactive television for voting and audience participation which raised the broad profile of SMS as a powerful platform for the broader consumer and small business market for Phase 3.

To enable business to exploit SMS functionality for what are termed 'vertical' *applications*[4], technology innovations have been developed to enable PC based SMS messaging rather than on mobiles and for converting Email to SMS and vice versa.

The key question for the industry in relation to SMS is whether a fourth phase is on the horizon in which SMS is just a 'messaging module' in a variety of new multimedia services types and what will be the its relationship with GPRS based messaging.

4. UNDERSTANDING USER INNOVATION

In the previous sections of this paper the concept that users of technology can reinterpret the use of that technology in a way that the industry did not anticipate has been developed with reference to the history of the short messaging service (SMS) on mobile phones. This, it is observed, spurred responses from the industry, albeit rather belatedly, at various stages in the evolution of SMS to develop and refine the technological and commercial means of providing that capability. In this way it is postulated that users are

[4] Vertical applications are sectors like education, reality TV etc

functioning as innovators - inspiring innovation, which leads to different approaches in marketing or the modification of the product[5].

The history of SMS suggests that this process happens at various phases in the evolution of the use of a technological platform, often, but not always[6], promoting an innovative technological response in one part of the industry or another. This process, it was observed, often required certain other preconditions to be present so did not necessarily happen in a predictable or uniform way. In this *spiral of innovation* the user is doing much more than just stimulating demand for a new service. They are actually suggesting a different *way* to use an existing technology platform thus creating new technology[7].

The history of SMS, interpreted within this conceptual framework, demonstrates the value of understanding *user inspired innovation* for those involved in designing, producing and marketing technology. It illustrates very well how the technology driven innovation process at any one point can fail to take up or identify a technology that is potentially valued and meaningful to a substantial body of users and that furthermore, conventional market forecasting methods can fail to identify the latent potential in the user population. We contend that this, in part, relates to the conventional way in which users are involved in product design and development and market research.

(Bar and Riis 2000) suggest that conventional industry-led technology design, development and market forecasting are overly reliant on *lead users.* They say such reliance reinforces the status quo in product development as lead users *"tend to be more sophisticated in their understanding of a technology so tend to suggest improvement in the technology itself"* (ibid). This focus on the high technology end of the market can result in the locking out of other consumers as the product development process "*tends to become a professionalised activity*" that relegates lay users to the role of a passive consumer (ibid). Furthermore this reliance can lead to '*unsatisfactory innovations"* and *"deprives producers of the insights lay users may have generated"*.(ibid). They explain lay users[8] included in the technology innovation process are less constrained by established practice and that "*ensuring that the broad lay base... gets access to advanced technology in the early phases of the innovation cycle allows the rate of experimentation at all levels to increase*" (ibid).

[5] The term *product* here is taken to include services

[6] This is where the *user led innovation* opportunity can be lost

[7] A broad definition of technology is used here to encompass any new way of doing something in this case on a telecommunications platform.

[8] Lay users are users without expertise or professional training in a particular field

This part of the paper will examine the role of lay users in user innovation with reference to three user research studies done by the authors (Coutts 1998, 2002, 2003). The primary objective of the studies was to gain insights into the likely adoption of mobile communications technologies of which users had no prior experience. The work has provided insights into the role users play in the innovation process by finding new ways of using the product (i.e. innovating).

From the work we have identified the necessary preconditions for fostering user innovation involving the wider lay user population. (Coutts 2002, 2003). We suggest that adoption of the technology is part of the innovation process as users have to first appropriate the technology before they can be creative with it – or innovate. Both processes are phenomenological and therefore the authors suggest a qualitative research approach is required which involves *"focusing on understanding the nature of the phenomena and their meaning, rather than the incidence*[9].

To understand this phenomenon and its application, a new research methodology is needed which is effective in elucidating responses from target user groups who have had little first hand experience or even awareness of the technological capabilities being explored.

From a range of qualitative research tools, the authors developed a user research methodology which through one encounter with lay users would create the necessary preconditions for embracing cutting edge communications technologies. These technologies were presented to the users in a *'hands on'* demonstration by the technologist in the research team. User participants would have to first reach a level of comfort, confidence and familiarity in order to adopt the proposed technological capabilities being presented to them. The research team were able to be reflective about their research assumptions and methodology. This was particularly important for the technologist who was in a sense acting as a proxy for supply industry. The researchers employed a data collection method that captured the subtle attitudinal changes and insights that the users were expressing through this experience and the analysis method allowed the 'meanings' to emerge from out of the data[10].

Adopting the paradigms of the social sciences the research team postulated that the research methodology had to have a number of elements. It had to have the capacity to engage users in an exploration of the technology in the context of their real life experiences and culture. Since adoption decisions by users (as with all human behaviour) do not reside in the rationale domain only, the researchers needed to also capture and

[9] Association of Qualitative Research (AQR) http://www.aqr.org.uk/glossary/index.shtml?qualmktres

[10] Grounded theory

understand the perceptions and misperceptions users held about communications technology. This was particularly relevant because the target user group were not technologically savvy and included those elements of the population of lay users who are technologically resistant. Though this was not deliberately prescribed, the selection criteria did seek users with only basic experience of mobile telephony and computer technology. Since the user participants had no experiential reference point, the enquiry method fostered the users' understanding of, and comfort with, the more advanced mobile technological concepts and capabilities. Finally the research approach provided a stimulus to visualising potential utility of these future technologies with which they had no prior experience. User innovation involves a paradigm shift for both the users and producers through an iterative process. The research team not only acted as observers and facilitators but also as participants in the process. The research activity therefore had to provide an effective means of facilitating a change process in a short time frame. The data collected had to reflect the subtle changes taking place and this achieved through the rich text of their conversations and actions.

Focus groups have the required characteristics (Gibbs 1997) and were used as the medium for interaction with users and with elements of ethnography[11], anthropology, action research, action learning and grounded theory analysis informing the design and conduct of the focus groups.

The research team needed to understand the nature of the users' daily lives, the culture in which they lived and its imperatives in order to understand the context in which the technology would be of value and the form that value – elements of anthropological enquiry. To understand the barriers and enablers to adoption and potential innovation, the researchers needed to 'hear' the values the users hold, the beliefs and perceptions the users have about the technology and themselves, their capabilities and their needs - the elements of ethnography. In effect, the focus group had to simulate an accelerated process of adoption and in doing so provide the impetus for innovation. To take the users through the process of change, the research team used elements of action learning (Yong and Pauleen 2004). A grounded theory approach was used for data analysis.

The results of each project provided insights, at a particular point in time in the evolution of mobile technologies, into how lay users can inspire innovation and coming, as it were, out of left field act as a driver within the innovation spiral. The first project (Coutts 1998) was specifically aimed at a significant market sector that had failed to embrace the opportunities presented by mobile technology at the time. In the second (Coutts 2002) and

[11] Ethnography produces in-depth understanding of real-world social processes investigating the relationship between beliefs and action in social situations.(Forsythe 2001)

third research projects (Coutts 2003) the focus was on users' responses to future capabilities of digital wireless data technologies.

In the first research project (Coutts 1998) the authors demonstrated user innovation at the end of the production cycle. They reported that their research question had been premised on the wrong assumption that the technology needed to be modified. Instead the products (value added mobile services) were found to be little used due to misperceptions held by the target group about how they might meet their service sector needs. At the beginning of the focus groups participants reported that they couldn't *see* themselves using mobile phones in their work. However, through examination of their organisational context a new value proposition was created. The study concluded that *"the marketing message of how to integrate wireless technology into modern life is not clear to many potential users"* (Coutts 1998) The outcome was a marketing strategy tailored to the cultural and structural needs of this sector – this was the user inspired innovation.

In their second research project (Coutts 2002) the authors refined the methodology to aid visualisation by users. This work was seminal in understanding of the *meaning* users give to the mobile phone which was akin to a form of technological embodiment or 'extension of self'. Because of the very personal relationship users have with their mobile phone (in contrast to their PC) the findings suggested that the impetus for innovative developments in m-commerce applications would come from mobile technology use rather than experience with e-commerce.

The third project (Coutts 2003) explored user reaction to emerging multimedia mobile services for lay users but was segmented on the basis of age, gender and socio-economic status The results indicated the basis for what we have termed a phase shift in the innovation spiral as the female users saw media (both stills and video) transforming the nature of the value proposition. However, the meaning or value proposition for the professional sector was around efficiency gains in essential communications and time management[12].

In all three studies, users reported that from initially seeing little value in the technology being presented they progressed to being able to envisage future use for the technology for themselves[13]. This phenomenon is a key element in the creative process that precedes innovation.

[12] Features including diary and contact lists for example that are synchronised with the office data.

[13] Note we would stress these were lay users, most of whom would not be regarded as early innovators in terms of technology adoption.

Our review of the three studies demonstrate that the research methodology was effective in revealing the nature of user innovation namely:

- innovation should be viewed as a phenomenon
- changed thinking is required for innovation
- old meanings held by users about the technology need to be understood and new meanings (or value propositions) have to be found
- understanding the necessary preconditions for innovation includes context of use and perceptions of value

The user research methodology simulated the adoption process which led to some users in the groups to appropriate the technology and so inspiring innovation. The presence of a technologist in the team of social researchers helped users to find value in the technology within their own context and translate this into potential product innovation.

This review of the three studies contributes to an understanding of the phenomenon of user innovation. To willing listeners it serves to inform those involved in technology design and development on how they can more easily capture the potential of this dynamic so that they are less likely to miss the opportunities presented by the 'SMSs' of the future.

5. DISCUSSION AND CONCLUSIONS

Technology innovation takes place through an evolutionary process involving a complex and interdependent interplay of product innovation and market diffusion. We have described a theoretical framework for considering this evolutionary process in distinct phases corresponding to the changes in the economic environment, market conditions and the state of the product. The argument developed in this paper is that new technologies or products that are spectacularly 'successful' are those that give rise to a spiral driven by the interaction of product innovation and market innovation in at least one of the phases of the product evolution.

This theoretical framework is explored in the context of one of the spectacularly successful new technologies, the SMS service product part of the GSM digital mobile technology. The case history identifies four distinct phases in the evolution of SMS from specification in phase 0, its modest beginnings in the market in phase 1 to the spectacular growth to a mass phenomenon in phase 2. In phase 3 where SMS is a commodity, it has now become a messaging building block for business messaging services. This study is a macro view of the phases of the innovation spiral throughout product evolution that demonstrated that there is a self-reinforcing process of product and user innovation.

A closer examination of the user's role in this innovation process was then made through appraisal of three 'user centred' research projects conducted by the authors over seven years. The research provides key insights into the lay users' perspectives and how their appropriation of technology to meet different uses inspires further innovation. All three projects looked at users and their views of mobile technologies including evolving digital services on the mobile platform. A modified focus group methodology was used to engage the lay user community in envisaging how they might use services of which they had no prior knowledge or experience. The research methodology set out was designed to take users through an accelerated process of adaptation and innovation at various phases in the evolution of mobile technologies.

While the user research projects were conducted at different points in time with to the respective product evolution history, they were commonly at the formative phases of the particular products. SMS was not specifically considered in the research as the focus was on emerging products. However, a change in user perspective regarding SMS consistent with the SMS history in Section 3 was observed between the two latter research projects. Such changes in user perspective over time could be effectively captured if incorporated in a longitudinal research process.

We recommend a longitudinal research process involving an annual engagement with a broad lay community of users to uncover the emerging phases of product adoption (and rejection). The focus group format could then be further refined to include part of the session to assess changes in user perception of 'current' products and a second part to consider new products. Such a two part methodology would enable the effective calibration of the phases of product evolution identified and better inform market forecasting, product development and marketing strategies.

We also suggest greater interaction of the research team in the observational room during the sessions using on-line discussion with the mediator and technologist conducting the focus group sessions would improve the user research methodology. While retaining the neutrality of the focus group facilitator, this potential for immediate interaction would enable the research to be more effective to enable testing of emerging insights while still retaining the rich records for reflective consideration.

The overall aim of the user research projects reviewed was to gain key insights into the likely take up of new technologies to enable conventional market research for demand forecasting. In this new innovation framework understands demand is understood in distinct phases that need to be identified. Therefore our future user research will develop the line of questioning for the focus groups and conduct the user sessions with this objective in mind. Further, the proposed more interactive involvement of the

research team with the focus group facilitator (and technologist) may enable the identification of what the key characteristics of potential different phases of market adoption of new service products might be.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support from the Smart Internet Cooperative Research Centre (www.smartinternet.com.au) for the research referenced in the paper up to the end of 2003 and the encouragement of colleagues to commercialize some potential products based on this innovation framework.

References

Bar and Riis (2000), Tapping User-Driven Innovation: A New Rationale for Universal Service, Information Society; April , Vol.16 Issue 2, p99, 10p

Christensen (1997), The Innovator's Dilemma: When Disruptive Technologies Cause Great Firms to Fail, Clayton Christensen, Harvard Business School Press.

Coutts (1998), 'A User Methodology – Identifying Telecommunications Needs', Pam Coutts, Communications Research Forum (CRF), Canberra, September 24-25.

Coutts, P. (2002) Banking on the move - Characterising user bottlenecks for m-commerce uptake, Communications Research Forum (CRF), Canberra, 2-3 October

Coutts. P., Alport, K., Coutts, R., and Morell, D. (2003),Beyond the Wireless Internet Hype – Re-engaging the User, Communications Research Forum, Canberra, 1-2 October

Coutts (2004), An Internet for Transactions: A User Perspective to Inform Future Business Models, Reg Coutts, Pacific Telecommunications Conference (PTC) Hawaii, January

Damsgaard and Gao (2004), A Framework for Analyzing Mobile Telecommunications Market Development, MOBIS'2005, 15-17 September, Oslo, Norway

Forsythe (2001), Studying Those Who Study Us An Anthropologist in the World of Artificial Intelligence, Diana E. Forsythe, Stanford University Press.

Gibbs, A.(1997), Focus groups", Social Research Update, Issue 19, University of Surrey.

Ito, M. and Okabe, D. (2003), Mobile Phones, Japanese Youth, and the Replacement of Social Contact, Unpublished.

Trosby, F. (2004), SMS, the strange duckling of GSM, Telektronikk 3.

Utterback, J. M. and Afuah, A. N. (1995), The Dynamic "Diamond": A Technological Innovation perspective, The International Centre for Research on the Management of Technology, MIT, November

Yong, P. and Pauleen, d. (2004), Generating and Analysing Data for Applied Research on Emerging Technologies: a Grounded Action Learning Approach, Information Research, Vol. 9 No. 4, July

THE EUROPEAN MOBILE DATA SERVICE DILEMMA

An empirical analysis on the barriers of implementing mobile data services

Martin Steinert and Stephanie Teufel

iimt (international institute of management in telecommunications),
University of Fribourg / Switzerland,
Boulevard de Pérolles 90, CH-1700 Fribourg, Switzerland. www.iimt.ch

Abstract: Based on a survey of 294 Swiss businesses, the study describes the desolate state of the art of mobile data services in relation to forecasts and different investment strategies employed by companies. Additionally, an investment confidence indicator is presented. In a next step, using a contingency analysis, it may be concluded that in Switzerland, which may act as an indictor country for Western Europe, there exists a (set of) common factor(s) of non-technical nature, which detain(s) companies from adopting mobile data services. A probing insight into the reasons why companies refrain from implementing mobile data services is given afterwards. Lastly, a concluding summary as well as a reference to a continuative research project is provided.

Keywords: empiric survey, Western Europe, mobile data services, barriers of implementations, contingency analysis.

1. THE PENDING SUCCESS OF MOBILE DATA SERVICES IN EUROPE

After the transformation of IT services by the global success of the internet and its web based services, the eyes of the ICT community have been focusing on equivalent developments in the mobile sector. Having overtaken the internet in terms of penetration the mobile hand sets in their various embodiments are already deeply integrated in the everyday private and business

life. In fact, GSM, the European second generation mobile communication system and its extensions are considered as "the fastest growing communication technology of all times" (GSM Association, 2005) with a global market share of roughly 75% out of a total subscriber base of ~1.700 million in terms of users in December 2004. (GSM Association, 2004) (EMC (European Mobile Communication) Database, 2005) Although this success is undisputed, the identification of mobile data services as the future sources of revenue and earnings growth has grossly misfired. In 2001, the idea was that mobile data services would given "Europe a crucial head start in the race toward development of a new era of wireless services that [would] make wireless Information Society a reality" (GSM Association - GSM Europe, 2001, p. 2). The EITO (European Information Technology Observatory) leaned even further out of the window. Besides forecasting an explosive growth in mobile E-commerce users, the EITO Report 2001 predicted, an European market Annual Revenue per User (ARPU) increase for mobile E-commerce from "less than Euro 200 in 2000 to about Euro 500 in 2005" (EITO (European Information and Technology Observatory), 2001, p. 276). One has to note that the forecasted 500 EUR comprises mobile data services only. Even in 2003, after the ICT bubble had defiantly burst, EITO went even further predicting a CAGR (Compound Annual Growth Rate) on mobile Commerce in Western Europe between 2002 and 2006 of 140% - annually notabene. (EITO (European Information and Technology Observatory), 2003, p. 40)

However, as anecdotal evidence strongly suggests, this bright picture of the future of mobile data service was too rosy by far. Nokia's Mikko J. Salminen realized in 2003 that "... we were sailing in tornado on the top of mobile hype curve" (Anonymous, 2003); he recognized a diffusion gap of the mobile future of around 5 years. Already in 2002, Lars Boman, the head of Ericsson Mobile Internet, had to admit: "The reality is that the introduction of new technology in networks and mobile devices has taken a longer time than anticipated. In addition it has taken a longer time for operators to turn the technology into compelling end-user services" (Timo Poropudas, 2002). The introduction of 2.5G technology such as HSCSD and GPRS but also EDGE and the rollout of UTRAN-FDD networks, often dearly paid for, has been continuously delayed. In private, executives admit that the cost of 3G networks have simply been written off – there is no intention left for 3G business case to break even, let alone to make a profit.

Following up on the pending success of mobile data services in Europe, this study explores the described dilemma empirically.

2. OBJECT, INTENTION AND COMPOSITION OF THE PAPER

It is the aim of this paper to firstly pinpoint the state of the art of European mobile data services and to secondly explore the question why businesses refrain from introducing and using mobile data services.

The mobile data services of interest in this project are packed switched services running over an air interface on the physical layer and using either next generation mobile communication networks (2.5G and 3G) or wireless internet (WLAN) networks as bearer service. Following Porter's argumentation (Porter, 2001), services are divided into two types: one concentrating on operational effectiveness and one on creating value directly for the customer. Depending on the distribution, different strategic approaches may be derived.

The observed business theatre is Switzerland, which may act as an indicator country for the Western European economies because of its well established ICT market and its proven test market capability.

A brief introduction of the methodology and the underlying data of the study (see 3.1) precedes the empirical observation on the 294 Swiss businesses in the sample. The intention of the study is to firstly provide a descriptive view on the current adoption of mobile data services as well as future concrete investment plans of such services (see 3.2). The results are to be compared with the 140% CAGR forecast of EITO in 2003. Next, a pattern of companies embracing or refraining from mobile data services will to be established via a cross table contingency analysis (see 3.3). The resulting picture is further elaborated by an explorative study of qualitative nature (see 3.4). Finally, an outlook is given, putting the results into an international perspective and recommending further research.

3. THE SURVEY: FUTURE OF MOBILE DATA SERVICES

This chapter presents the result of an empirical survey of Swiss Businesses on their deployment of Mobile Data Service. After giving some general information about the underlying sample and methodology the descriptive results will be presented: current usage of Mobile services and planned investment in Mobile services including a investment confidence indicator. Next the results will be check for statistical associations followed by an explorative empirical analysis on possible barriers to implement Mobile services.

3.1 Basic Information on the Survey

The data presented in this paper originates from the telecom guide Switzerland, an annual survey of Swiss Businesses by the international institute of management in telecommunications (iimt) of the University of Fribourg / Switzerland. The telecom guide Switzerland, published annually since 1998 targets the CTOs (Chief Technology Officers) of Swiss businesses. It aims to continuously track the subjective satisfaction of businesses with their ICT provider in terms of fix line services, mobile services and ISP services. "Future of Mobile Data Services", the add-on to the 2003 questionnaire, out of which the results are being presented in this paper, starts with a brief introduction including a definition of the services and mobile technologies in question. The next three question sets (nominal values) had to be answered by the respondents. An open, qualitative question concludes the questionnaire's mobile service special.

Due to the B2B nature of the survey, the universe of interest is composed of Swiss businesses. Although, in 2001, there have been 317.739 business registered, which form the theoretical universe; of theses, only about 52.000 qualify as active in the B2B market. (Bundesamt für Statistik (BFS), 2002, p. 4) (Schlienger *et al.*, 2003, p. 7) They have been selected as basic population of the study. Those 52.000 companies are enclosed in the database from Kompass. (Anonymous, 2005). Out of the data sets in the database, 32.755 qualified for further use (the other data sets were incomplete in terms of sector affiliation and/or company size. Out of the remaining data sets, a random sample of 5.000 companies has been selected. They were contacted in autumn 2003 via mail. An additional 5.000 companies have been addressed by Euroforum, the sponsor and media partner of the telecom guide Schweiz 2003. (Euroforum, 2005).

The special on mobile data services has been answered by 294 companies which equals a return quote of 2.87%. (n=294) Of theses companies, 22% chose to answer online the other returned the filled out questionnaire by snail mail or fax. For a comparison in terms of internet proficiency of Swiss businesses, the OECD has published the following numbers for 2002: PCs/Workstations/Terminals diffusion 96.2%, E-mail usage 92.7% and internet usage 92.1%. Nevertheless, most businesses chose to traditionally answer the physical questionnaire. (OECD - Organisation for Economic Co-operation and Development, 2004, p. 9)

Concerning the sample allocation in terms of cultural regions, the distribution of languages in the sample was as follows: German speaking 84% and French speaking 16%. This compares to the language distribution in Switzerland: German speaking 64%, French speaking 20%, Italian speaking 6% Romantsch speaking 1% and 9% other languages. It may be noted, that

the sample represents the general Swiss allocation between the German and the French speaking population – though with a slight prevalence of German. The excess weight, however, is justifiable on the basis of the higher share of economic activities of the German speaking cantons in Switzerland.

In terms of branch affiliation and company sizes, the sample was categorized based on the Swiss Federal Statistical Office official classification in categories: Nomenclature Générale des Activités Economiques (NOGA), which equals the general classification for economic activities in the European Community (NACE). The production and trading sector accounts for 58% in Switzerland (excluding agriculture) compared to the 34% in the sample. (Schlienger et al., 2003, p. 14) The service sector (excluding agriculture) accounts for 42% in Switzerland, compared to 51% in the sample. (Schlienger et al., 2003, p. 14) It can be seen that, in contrast to the actual Swiss statistics, companies active in the service sector have a higher share in the sample. Concerning the sample distribution on sizes, small and medium companies are defined according to the Swiss federal statistics as having less than 250 employees. In the sample, big companies are, with 51%, heavily overrepresented. SMEs' comprise 99.7% of the legal Swiss business entities and employ 67% of the national workforce.

Methodologically, the survey has been constructed as a mail survey although is has also been possible for respondents to answer the questionnaire online. For the mobile service special two types of services have been differentiated: (Schlienger et al., 2003, p. 63-64)

Mobile services, which improve operational effectiveness (Type I). Theses services aim to advance the company's production process, ultimately, by reducing the costs inside the value chain. Any mobile service reducing the time or the costs to produce a certain physical output or intangible service falls into this category. Type I services improve the effectiveness of current activities.

Mobile services, which interact directly with outside customers (Type II). Services of this type may take the form of an entirely new service, or a supplementary service to create value-add to existing products and services. Basically all company mobile services, which generate a new, direct contact to customers fall into this category. Type II services consequently approach new customers in a new way, or known clients via a new or partial new service.

Secondly, the mobile services were differentiated via their base technology: (Schlienger et al., 2003, p. 64)

- GSM (SMS) (Circuit Switched Mobile Telephony Data Services)
- UMTS/GPRS (Packet Oriented Mobile Telephony Data Services)

- WLAN (Wireless Local Area Network based on IEEE 802.11)
- Bluetooth (Short-range radio signal - wireless networking for devices)
- GPS (Satellite based Location Information)

The nomenclature has been kept simple in order to increase the survey return. The questions asked aimed for the current usage of mobile services and the planned investments in mobile services; its results will be presented next.

3.2 Descriptive Results

The descriptive result of the study will be presented next. In the first part of the survey, Swiss businesses were asked to state whether they were currently using Mobile data services. An answer has been demand for each of the five technologies mentioned above. The data from the 294 businesses are converted to percentages and stated separately for services Type I (operational effectiveness) and services Type II (customer interaction).The descriptive results are presented in figure one; table one contains the according values in percent.

The data for Services Type I, which increase the internal operative effectiveness, suggests that only about 20-25% of Swiss businesses are using some kind of Mobile Date Service to decrease their internal costs. The obvious exceptions are GSM based services which are used by up to 60% of the businesses. A less fortunate picture presents itself when the same question is asked for services Type II which create direct customer interactions. It is obvious, that Mobile services Type II, i.e. services which create a new or improve an existing contact with a customer, are much less en vogue. GSM based services are still fairly in use with up to 40% of the businesses but services based on other technologies are only deployed in 4-13% of the questioned companies. More intriguingly, Type II services only are deployed by less than 3% of the companies (GSM is the exception with 5.8%).

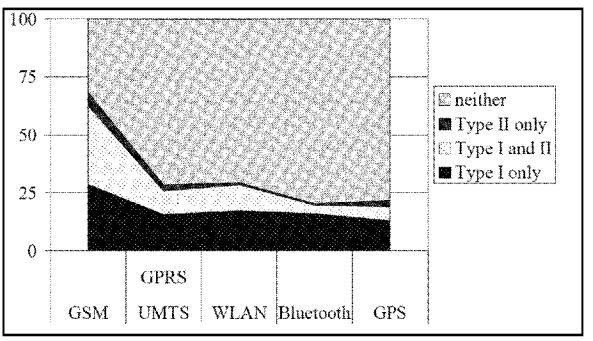


Figure 1. Current Usage of Mobile Services in percent

Table 1. Current Usage of Mobile Services in percent (n=294)

	GSM	UMTS GPRS	WLAN	Bluetooth	GPS
Type I only	28.6	15.6	17.3	16.0	12.9
Type I and II	34.0	10.2	11.2	3.7	6.1
Type II only	5.8	2.7	1.0	0.7	2.7
Neither	31.6	71.4	70.4	79.6	78.2

As a first result it must be noted that about a quarter of all companies have introduced some kind of mobile service. Most of these services target internal processes, i.e. cost cutting or profit maximization. Services which may create new communication and distribution channels to the customers (or improve the existing ones), i.e. services which may increase future revenue or Type II service are deployed only by a fraction of companies, which have opted for services of Type I. This result suggests that the deployment of mobile services originates of the internal resources strategy of companies, rather than their market development strategy. Before questioning the results more closely, the investment mood in mobile services will be presented.

Businesses were asked whether or not they intent to further invest in Mobile services during the next fiscal year. This way, the survey intents to obtain time-series relevant information, without having to be conducted more than once. Instead or solemnly stating one point in time, a more robust trend can be elaborated from the data. Parallel to the last figure and table, the results are presented in figure two and table two.

Concerning future investments, the results allude to a more uncomfortable investment climate. Further investments in WLAN based services are supported by only around 30% of the companies. For all other services the rate decreases to between 15% and 20%. When comparing the results with figure four, it is apparent that investments are actually slowing down in comparison to services in use. Still, Type I services seem to be favored over Type II services. Therefore, the near future should not lead to an increase of mobile services Type II relative to services Type I. The customer is not "yet" in the cross hairs of the companies.

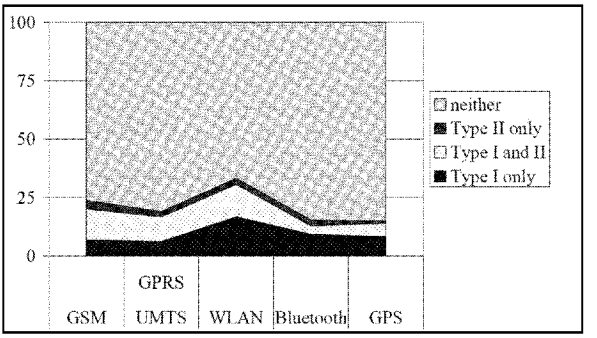


Figure 2. Companies planning Further Investments in Mobile Services in percent

Table 2. Companies planning Further Investments in Mobile Services (n=294)

	GSM	UMTS GPRS	WLAN	Bluetooth	GPS
Type I only	6.8	6.1	16.7	9.2	8.2
Type I and II	13.3	10.5	13.9	3.7	5.8
Type II only	3.7	2.4	2.7	2.7	1.0
neither	76.2	81.0	66.7	84.7	85.0

To get a better picture on the investment behavior and especially in the confidence in the investments, an indicator is presented next. Companies were asked, if those investment plans had been revised up- or downwards, in relation to the current fiscal year. The aim of this question is to create a trend indicator for investments in mobile services in addition to the current investment situation presented. This question turned out to be a little delicate with respect to the confidentiality of internal decisions. Accordingly, only some of the 294 companies chose to answer (58-87, depending on the mobile service in question). Thus, a different way to analyze the data has been chosen. Instead of giving a percentage, which would reflect the whole sample, a trend indicator is shown, based on summing up the positive (dec_{pos}) or negative (dec_{neg}) decisions to revise the current investment plans.

The resulting confidence indicator (con_ind) is calculated in the following way:

$$con\_ind = \frac{\sum dec_{pos} - \sum dec_{neg}}{\sum dec_{pos} + \sum dec_{neg}}$$

con_ind = confidence indicator

dec_{pos} = positive investment reassessment decision

dec_{neg} = negative investment reassessment decision

The confidence indicator results in values, ranging from +1, which equals unrestricted upward revision of the investments of all answering companies, and -1, indicating that all companies have revised their investment decision downward. Table three present the results.

Table 3. Confidence Indicator (n=294)

	GSM	UMTS GPRS	WLAN	Bluetooth	GPS
dec_{pos}	30	24	44	14	18
dec_{neg}	51	39	43	44	41
missing	213	231	207	236	235
con_ind	-0.26	-0.24	0.01	-0.52	-0.39

The investment confidence indicator clearly points into a negative direction. The current investment mood is abating in the case of four technologies and merely neutral for WLAN. The already negative investments trend described in 2.2.2 is hence further supported. Since most companies focus on mobile services Type I, it seems that companies may develop a growing distrust in the cost-cutting abilities of mobile data services. According to the confidence indicator, only the further adoptions of WLAN seem to realize the promised benefits. A boom though is nowhere in sight.

When comparing the ambitious forecast of 140% annual growth for mobile commerce between 2002 and 2006 made by EITO (see page 2) and the results from the survey as presented in the last few sections of this papers the positive forecasts have to be seriously doubted. An extraordinary positive change in demand and business activities is needed to reach the fare stretched goal -- as the investments confidence indicator implies though, three digit growth is unrealistic. Basically, for a company to achieve a lasting competitive advantage, Mobile data services can have two directions of impact: cost cutting or product/service differentiation. (Porter, 2001) Type I services support a cost-leadership or price strategy. Unfortunately, advantages for this strategy, as developments in the internet show, are short lived. After a short time span, competitors are usually able to obtain the same services, thus rendering the advantage useless. The achieved higher level of competitiveness by the innovation leading company will in due time diffuse into the entire industry. Eventually, the cost-cutting technology will become a standard technology and more of a "must have" than a distinguishing factor creating a competitive advantage. Market growth for Mobile services of this Type I services are based on clearly communicated cost savings, e.g. by installing a WLAN instead of laying cable in new buildings. However, as this study shows, the momentum for increased investments in type I services

is fading; with the exception of WLAN companies seem to think little more efficiency gains are to be expected by mobile data services. Type II services, on the other hand, are services which create new products in itself or which upgrade a product by offering a new distribution channel or by further improving existing products and services with added value. This Type II service strategy leads to a product/service differentiation for which the successful company may demand and get a price premium. After the prevailing opinion of business scholars, the successful introduction for new products or distribution channels also leads to a longer lasting competitive advantage. This contrasts to gains from a price strategy (Type I services). A differentiation strategy (Type II services) is usually more stable and enduring. Competitors have a harder time copying the advantage since the introduction of new services and the creation of new products usually take more time and investments than to simply introduce cost-cutting technology off the shelve. Instead a Type II mobile service strategy may even lead to the creation of entire new market niches, which, in time, may be able to challenge established industries due to their novel approach and superior technology. As such Amazon, Ebay and Google are probably the best current examples from the internet nexus. By creating niche markets based on technological advantage or customer satisfaction, Type II strategies tend to last and are often the first step to attack the incumbent later on from higher ground. However, in order to implement such success stories, companies need a strategic intent to discover and create the new value chain down to the customer. This intent must be sufficiently strong and determined to support investments which will not amortize in a single fiscal period. As the presented survey shows the intention to introduce Type II mobile data service seems to be limited to only a few companies. The disruptive evolutions needed to achieve the forecasted 140% annual growth in the industry seems to be far away. Having established this picture from the data, the question at hand is to ask why companies refrain from investing into Type II mobile services. Why has the praised mobile revolution, creating a vast amount of new application, not been launched yet? Approaching this question on the foundation of the empirical data, a contingency analysis is presented next.

3.3 Cross Tabulation and Contingency Analysis

In order to understand which factors are actually hindering mobile growth the empirical data base of the survey is further analyzed. The following theoretical background on cross tabulations and contingency analyses are mainly based on (Backhaus *et al.*, 2003, p. 229-257).

Since we are dealing with nominal variables a cross tabulation and contingency analysis may be used to expose and check hidden associations and

interrelationships. The aim is to uncover whether an association in the sample is based on coincidence or a systematic interrelationship. The data sets in question are the responses on the question on usage and on investment of mobile services. The resulting table spans a n_{ij} (i=5; GSM, UMTS/GPRS, WLAN, Bluetooth and GPS) (j=4; Type I only, Type I and II, Type II only, neither). Since there are only very few counts for Type II only, the data has been recoded for j=2 (mobile service, no mobile service) in order to prevent insufficient cell counts which would result in meaningless results.

The contingency analysis now statistically checks the independence of the attributes combined in a cross table; in our case it compares the companies' responses concerning their mobile data service usage/investments per technology. For example the usage of GPS based services vs. the usage of Bluetooth based services. The statistical hypothesis H_0 is: X and Y are independent from each other, i.e. the decision of the companies to use GPS based services is independent from the decision to use Bluetooth based mobile services. The statistical test generally used is Pearson's Chi-Square χ^2-test. However, if the cell count is between 20 and 60 items, Yate's Continuity Correction has to be calculated; for cell counts below 20, Fisher's Exact Test is used. Cell counts below 5 produce meaningless results – there are no such events in the presented analysis. WLAN x GSM has the smallest cell count in the table with 18; Fisher's Exact Test applies.

The second important question after uncovering a dependency is to measure the strength of the interrelationship. An often used measurement is the Phi Coefficient (φ) and Cramer's V. Since both are identical if binary variables are used (j=2), the values are presented only once. φ may result in values above 1; in this case literature suggests using the Contingency Coefficient. This case does not apply. Table four presents the results of the contingency analysis on the basis of the data concerning the usage of mobile services.

Table 4. Contingency Analysis - Usage of Mobile Services (n=294)

	UMTS GPRS	WLAN	Bluetooth	GPS
GSM	.000*** 3) .268**** 4)	.009*** 3) .153**** 4)	.000*** 3) .236**** 4)	.000*** 3) .252**** 4)
UMTS GPRS		.000** 3) .332**** 5)	.000*** 3) .446**** 5)	.000** 3) .250**** 4)
WLAN			.000*** 3) .300**** 4)	.428** 1) .055**** 4)
Bluetooth				.000** 3) .244**** 4)

The significance tests are labelled as follows:

* Pearson's Chi-Square, Asymp. Sig. (2-sided) / cell size >60

**	Yate's Continuity Correction, Asymp. Sig. (2-sided) / cell size 20-60
***	Fisher's Exact Test, Exact Sig. (2-sided) / cell size 5-19
1) > 0.05	(not significant association)
2) < 0.05	(significant association)
3) < 0.01	(highly significant association)

The strength indicator Phi is labelled as follows:

****	Phi Coefficient, Value (equals Cramer's V)
4) -0.3 to +0.3	(trivial dependency)
5) +0.3 to +1.0	(non-trivial, positive dependency)
6) -1.0 to -0.3	(non-trivial, negative dependency)

The surprising result presents highly significant associations over all technologies, with the exception of WLAN x GSM). In terms of strength, the dependencies are positive but indistinct though they exist with φ values around +0.025 to +0.30. To check the results a comparative contingency analysis of the same data but concerning the variables langue, company size, company branch affiliation and answer channel used has been calculated – no similar dependencies are apparent.

To check further, the same analysis has been run on the data collected concerning further investments planned in mobile services. Table five presents the results.

Table 5. Contingency Analysis - Further Investment in Mobile Services (n=294)

	UMTS GPRS	WLAN	Bluetooth	GPS
GSM	.001** 3) .197**** 4)	.000** 3) .282**** 4)	.057*** 1) .117**** 4)	.000** 3) .280**** 4)
UMTS GPRS		.000*** 3) .429**** 5)	.000*** 3) .227**** 4)	.000** 3) .355**** 5)
WLAN			.000*** 3) .261**** 4)	.002** 3) .189**** 4)
Bluetooth				.023*** 2) .139**** 4)

The significance tests are labelled as follows:

*	Pearson's Chi-Square, Asymp. Sig. (2-sided) / cell size >60
**	Yate's Continuity Correction, Asymp. Sig. (2-sided) / cell size 20-60
***	Fisher's Exact Test, Exact Sig. (2-sided) / cell size 5-19
1) > 0.05	(not significant association)
2) < 0.05	(significant association)
3) < 0.01	(highly significant association)

The strength indicator Phi is labelled as follows:

****	Phi Coefficient, Value (equals Cramer's V)
4) -0.3 to +0.3	(trivial dependency)
5) +0.3 to +1.0	(non-trivial, positive dependency)
6) -1.0 to -0.3	(non-trivial, negative dependency)

Again, all five technological investment decisions are significantly inter-dependent (with the sole exception of GSM x Bluethooth) – again the strength of the dependency is rather weak but aligned positively.

Before drawing conclusions, it must be noted though that the contingency analysis is not able to verify cause-effect relationships but correlations only. Concerning a possible bias it must be noted that the presented result originate from one single survey sample, that is from one point in time. The construct validity and reliability though can be considered as high.

Concerning the interpretation of the analysis, the following may be stated: Based on the cross tabulation and contingency analysis, the decision of the 294 surveyed Swiss companies whether to use mobile data services or not and whether to further invest in mobile data services or not, correlates across mobile data service of all five technologies, GSM, UMTS/GPRS, WLAN, Bluetooth and GPS. A (set of) common factor(s) detains companies from adopting mobile data services or support companies to adopt the same.

Concerning the mobile data services in question, the five scanned technologies are disparate on a multitude of technical attributes, e.g. concerning: frequency, bandwidth, transmission rate, range , two-way communication capability, possible reception speed, localization capability, end device and its features, regulation, standardization, market diffusion, life-cycle, service price, target audience etc.

The conclusion may be drawn that there are no significant technological factors relating all five technologies. A thorough discussion of this aspect as well as an extensive argumentation can be found in (Steinert, forthcoming). Consequently, the common factor mentioned is of non-technical nature.

Summed up, the contingency analysis of the cross tabulation imply, that: *"In western Europe, there exists a (set of) common factor(s) of non-technical nature, which detains companies from adopting mobile data services!"* In order to approach the questions on what kind of factors may influence the companies' decision, the results of open qualitative question of the same study are presented next.

3.4 Explorative Results – Barriers of Implementation

In the questionnaire, the Swiss companies were also asked in a qualitative open question to state from their perspective, which reason might hinder the implementation of mobile data services. Multiple answers up to three per questionnaire were allowed. 157 companies answered this question and gave a total of 318 reasons. Methodologically, 11 clusters of possible reasons have been nominated by the author. Afterwards, two researchers have independently put the 318 reasons into the 11 answer clusters. In case of diverging allocations, a third researcher has decided independently. The following 318 reasons present the resulting barriers of the implementation of mobile services:

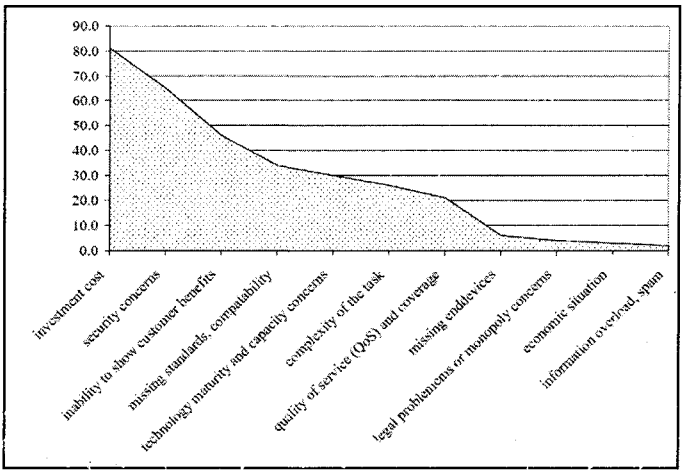


Figure 3. Barriers of Implementation of Mobile Data Services in counts

Table 6. Barriers of Implementation of Mobile Data Services

Group	Answer cluster	Counts
1	investment cost	81
1	security concerns	65
1	inability to show customer benefits	46
2	missing standards, compatibility	34
2	technology maturity and capacity concerns	30
3	complexity of the task	26
2	quality of service (QoS) and coverage	21
2	missing end devices	6
3	legal problems or monopoly concerns	4
3	economic situation	3
3	information overload, spam	2

Interestingly, again, reasons of technical nature do not seem to be at the forefront of companies' concerns. Instead, investment costs, general (non technical) security concerns as well as the inability to identify and show customer benefits are forming the first group of reasons. This group account for ~60% of the given reasons. The second group of concern which accounts for ~28% of the given answers is formed by technical reasons such as missing standards and missing compatibility, technology maturity and capacity concerns, Quality of Service (QoS) and coverage concerns as well as missing end devices. The last group which accounts for the remaining ~9% of the answers, includes reasons such as the complexity of the implementation task and minor reasons such as legal problems or monopoly concerns, the current economic situation or fear from information overload or spam. Of course, this last exploratory analysis is highly biased but it points in the same general direction as the contingency analysis. Summed up, in their decision to not yet adopt mobile data services, Swiss companies are not concerned with roll out problems or power supply concerns. Instead the underlying business case offered by the providers is not convincing enough to make them invest into mobile data services.

4. CONCLUSION AND OUTLOOK

The paper at hand is founded on a survey of 294 Swiss businesses concerning their usage and investment plans of mobile data services.

In a first step, it analysis the state of the art of mobile data services in Europe and compares it with the farfetched growth assumptions of e.g. EITO. Reviewing the empirical data, instead of the predicted 140% CAGR, a boom of mobile data services is nowhere in sight. Companies have so far concentrated their activities on internal services in order to cut costs. New customer relationships or the added value of existing customer relationships and hence a possibly lasting competitive advantage seem to be unattractive. For the worse, the investments confidence of companies is decreasing as well.

Based on contingency analyses, it may be assumed that across different technology platforms including GSM, UMTS/GPRS, WLAN, Bluetooth and GPS, there is a (set of) common factor(s) of non-technical nature, which detains companies from adopting mobile data services.

A first probing via an open qualitative questions results in the missing business case as a possible argument against implementing mobile data services. Providers do not seem to offer a convincing investment story (cots vs. benefits) to persuade customers to invest in mobile data services. Legal and general economic concerns do not seem to matter.

The argument scheme of this study is the starting point of a current research project to further probe into the reason why European mobile data services fail to keep up with promises and forecasts as well as with the successful past of GSM voice services. In a comparative case analysis the macro and micro economic as well as technological and legal developments in Western Europe are compared with the two successful Asian mobile powerhouses: South Korea and Japan.

References

Anonymous. (2003). Year 2000 comes in 2006, *Mobile Monday*.

Anonymous. (2005). Kompassonline.Ch: Kompass.

Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2003). *Multivariate analysemethoden* (10 ed.). Berlin Heidelberg New York: Springer.

Bundesamt für Statistik (BFS). (2002). *Unternehmen, arbeitsstätten, beschäftigte - die betriebszählung 2001 in kürze*. Neuchâtel: Bundesamt für Statistik (BFS).

EITO (European Information and Technology Observatory). (2001). *Eito report 2001*.

EITO (European Information and Technology Observatory). (2003). *Eito report 2003*.

EMC (European Mobile Communication) Database. (2005). World cellular information service. Retrieved 09.05., 2005, from http://www.emc-database.com/NASApp/cs/ContentServer?pagename=marlin/home&MarlinViewType=MARKT_EFFORT&marketingid=20001226000&siteid=30000000401

Euroforum. (2005). Über euroforum. Retrieved 10.05., 2005, from http://www.euroforum.ch/about/

GSM Association - GSM Europe. (2001). Facts and figures - about gsm europe. Retrieved 09.05., 2005, from http://www.gsmworld.com/gsmeurope/about/gsm_europe_factsfigures.pdf

GSM Association. (2004). Growth of the global digital mobile market. Retrieved 09.05., 2005, from http://www.gsmworld.com/news/statistics/pdf/gsma_stats_q4_04.pdf

GSM Association. (2005). Gsm facts and figures. Retrieved 06.05., 2005, from http://www.gsmworld.com/news/statistics/index.shtml

OECD - Organisation for Economic Co-operation and Development. (2004). Ict diffusion to business: Peer review, country report: Switzerland. Retrieved 01.05., 2005, from http://www.oecd.org/dataoecd/28/12/31706392.pdf

Porter, M. E. (2001). Strategy and the internet. *Harvard Business Review, 79*(3), 62 (17p).

Schlienger, T., Steinert, M., & Unterberger, C. (2003). *Iimt telecom guide schweiz*. Fribourg: iimt University Press.

Steinert, M. (forthcomming). *A comparative cross case analysis on western european and asian mobile data services*. Unversity of Fribourg / Switzerland, Fribourg.

Timo Poropudas. (2002). Slow down to reality, mobile hype is over, *Mobile CommerceNet*: CommerceNet Skandinavia.

ON THE DEVELOPMENT OF AN OPEN PLATFORM FOR M-GOVERNMENT SERVICES

Helena Rodrigues, César Ariza and Jason Pascoe
Dep. Sistemas de Informação, Universidade do Minho, 4800-058 Guimarães, Portugal

Abstract: Citizens and local government services are often not as well connected as they ideally should be. Services may not be well advertised or may simply be cumbersome or time-consuming to access. In this paper we present our ongoing work in investigating how to better connect the citizen with their local government with the support of context-aware applications. We describe a set of user requirements for m-government services and open service-oriented platforms. In particular, we analyze the requirements and present the research issues on a context modeling component for supporting context-aware service discovery. Our motivation to develop it is driven by the need to provide appropriate local government services in an easily accessible manner whenever and wherever they may be needed.

Keywords: m-government services, m-government user requirements, context-awareness, context-aware service discovery.

1. INTRODUCTION

Mobile technologies provide an important alternative channel for the interaction between authorities and citizens. This channel is very important as such for two reasons. Firstly, they are accessible to many more people than traditional desktop computing. The penetration rate for mobile phones in most European countries is already above 75%, allowing services to be reached by a much larger population and helping to break the digital divide. Secondly, these channels have the characteristics of allowing nearly immediate communication, thus supporting time-constrained notifications that would not be possible by other channels, and allowing applications such as alerts, traffic reports, complaints and promotion of events to be implemented.

Our work is part of a pan-European project called USE-ME.GOV (USE-ME.GOV consortium, 2002) whose general objective is to promote a better connection between citizens and their local government services through mobile technology. The project vision is centered on the provision of *appropriate* services (i.e context-dependent services) to the citizen directly in the time and place in which they are needed and/or useful. Context-awareness (Schilit et al., 1994), the ability of a system to sense, react, and adapt to its environment of use, can be used to discover, filter and prioritize the set of services appropriate for a citizen's current context. For example, considering a service to report anomalies to public authorities, when located in a park it is quite probable that the citizen's complaint will be directed to the parks and gardens authority, a nearby a water main may also present a possible source of concern in the event of a burst pipe, or a history of graffiti in the area may prompt the system to select the graffiti correction department, and there is also a good chance that the citizen's complaint may be the same as other recently reported complaints in the locality.

There have been recently several approaches to m-government (see section 2). The reviewed projects mainly present the lack of support from an open and interoperable architecture for the sharing of external content providers, mobile operator interfaces, context information and third-party services. In this paper we present our work that consisted in the analysis of a set of user requirements for m-government services and the derivation on a set of systems requirements for an open architecture for the development of context-aware applications. We mainly focus the requirements for a context modeling component as the support for m-government context-aware applications.

The paper is organized as following. In section 2 we review existing m-government platforms. In section 3 we describe the set of user requirements for an open service platform for m-government services. In section 4 we describe the logical architectural model of USE-ME.GOV platform and describe, particularly, the requirements on service repository and service discovery components, as well as requirements on context modeling component. In section 5 we present our work on a context modeling component design and, finally, in section 6 we present our conclusions and future work.

2. REVIEW OF M-GOVERNMENT PLATFORMS

In the last years there have been several approaches to mobile government. Such approaches share common objectives with USE-ME.GOV, although addressing different issues. The mGovernment Service

initiative in Malta (Government of Malta, 2003) offers e-government services to mobile users using SMS (Short Message System) technology. Services are limited to notification services as the interaction format is mainly text and are not context-based. As far we understood, the components of the technological architecture are mainly the government services and a component that communicates with mobile operator for routing the SMS messages. The main limitations of this architecture is that it does not offer any API (Application Programming Interface) for sharing content between services, for integrating third party service providers or for integrating new mobile applications.

The Visual Admin project initiative (Visual Admin consortium, 2000) offers significant functionality to citizens through an e-government Web Portal. The ubiquitous access they provide is limited to offering a new channel to access the portal (GPRS), giving up all the opportunities offered by the mobile communications technology such as the location services components. Actually, one of the results of the project states that users mentioned that some of the functionalities on the fixed interface should not appear in the mobile one. The technological architecture is mainly based on database and search engine technology, what may compromise scalability and openness.

The Pandora project (Pandora consortium, 2001) presents some similar objectives to the ones of USE-ME.GOV in what concerns the provided functionality to the user, although not considering context-awareness. The project objective is to specify and implement a new Mobile Content Management Platform able to access, manage and deliver wireless multimedia/multilingual information from Web pages, local/regional databases, and other content sources, exploiting UMTS technology features. At the current stage, the main objective of USE-ME.GOV project is to specify and implement a platform for the deployment of e-government services, evolving from the traditional approaches of Web information filtering to the concept of an open platform that offers sharing of common content and application programming interfaces for the development of new services and mobile applications.

The CENTURi21 project (CENTURi21 consortium, 1999) also presents some similar objectives to the ones of USE-ME.GOV in what concerns the provided functionality to the user, although not considering context-awareness. The results on interoperability and integration of services may be of USE-ME.GOV interest. However, the target context of USE-ME.GOV project is a local context as opposing to a global target context of CENTURi21 project. The late one does not explores the relevance of user context and location to improve both usability and sharing.

The Mobhaile project (Mobhaile project, 2004) presents a set of very similar functionalities to the one in USE-ME.GOV, including context-aware functionality. Apparently, the supporting platform does not support integration of third-party services and application providers, compromising scalability and openness. Although offering a friendly interface, such as a Map interface, this platform does not apparently offer the advantage of mobile communications technology such as location services.

3. ANALYSIS OF REQUIREMENTS FOR M-GOVERNMENT SERVICES

3.1 Use Case Description

Functional requirements capture the intended behavior of the system from the users' point of view. This behavior may be expressed as services or functionalities the system is required to perform. In this section, we define a set of use cases built after an analysis of user requirements for m-government services in the context of USE-ME.GOV project. These use cases describe the set of functional requirements of USE-ME.GOV platform. We have defined two orthogonal criteria to breakdown the USE-ME.GOV platform functionalities (see figure 1) . The first one, criteria a (left), describes the main functionalities of the system and the second one, criteria b (right), defines the different application domains.

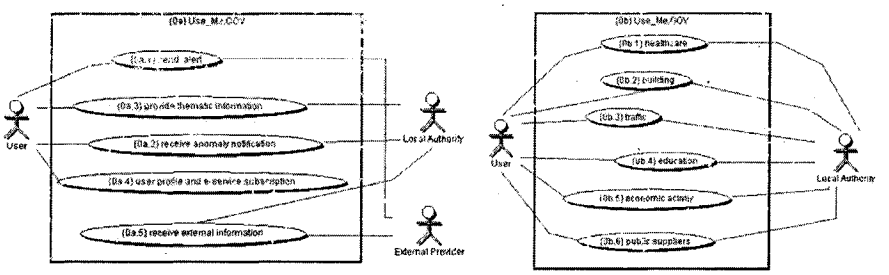


Figure 1. Functionalities of the m-government platform.

The user is the beneficiary of functionality provided by an application that is, to some extent, based on the capabilities of the USE-ME.GOV platform. We assume that the user is a subscriber of a mobile network and carries some communicating device.

The USE-ME.GOV platform assumes a special case of external provider that is the Local Authority. At this stage, we distinguish Local authority

from other external providers as they are the entities defining and providing the ultimate services to the user and are also a beneficiary of the platform.

The USE-ME.GOV platform assumes that many other entities can contribute to the system by providing general information (geographic information, traffic related information, tourism information ...) and mobile operators network services (the mobile networks themselves).

We are now presenting a discussion on each of the identified functionalities for an m-government platform in which are based our system's requirements.

send alert: This is an asynchronous service and refers to the dissemination of particular domain dynamic information to mobile users informing on specific events and situations or unexpected situations that are happening in the region. Such functionality is particularly interesting for the mobile setting as user context, such as user location, time, activity, preferences, etc, may determine the set of users the alert should be sent and determine the adequate granularity of information. For example, an alert service could be used for warning users in a certain geographic area about a traffic congestion, flood, weather storm, etc. The information associated to the alert should always be up-to-date and match the user-specific request, excluding any extra information or undesired advertisements. For those users that require personalized information, subscription services are necessary.

receive anomaly notification: This is a synchronous service that consists on receiving a report from users informing of some anomaly of a particular domain. Such functionality is particularly interesting for the mobile setting as this service may take into consideration the information on user context, such as location, to characterize the reported situation. For example, a user report, when located in a garden, would be direct to gardens authority. It would also be helpful to allow the user to navigate to the appropriate level of context information granularity, starting from the high-level context up to the level of user awareness, in order to get more accurate information on user context. For example, if the user chooses to complaint about the bad state of a flower bed in a garden, the dialogue may provide a set of gardens located in an automatic retrieved wide area location, and ask the use to select the correct one.

provide thematic information: This is a synchronous service that consists on provide information of a particular domain to user on demand. Such functionality is particularly interesting for the mobile setting as this service may take into account the user context information (user's geographic location, time, user's current activity) to select the appropriate information from the distributed information space. For example, if the user requests information on local events or local businesses, for example within the context of industrial or thematic fairs, the service may use his location

information to select the appropriate services. For those users that require personalized information, subscription services are necessary.

user profile and e-service subscription: The user registers into the system providing his profile and the set of subscribed services. Particularly, for alert functionality, users subscribe required alert services, for example. If users want to receive personalized information via both send alert and provide thematic information they should be able to create a user profile with personal identification and preferences specific to each particular domain service.

receive external information: The USE-ME.GOV platform includes the functionality to receive/request information in a particular domain from/to external content providers. For example, a geographic information system provides landmarks geographic positioning for supporting users localization process, or, a traffic information external provider provides traffic congestion alerts that may support traffic alert services. Contents/events and related information are to be delivered by the system (via send alert and provide thematic information) to the users basing on users' profiles and users' context.

The functionalities we have described exhibit a set of common characteristics. They characterize mainly a spontaneous user interaction with the system, the USE-ME.GOV platform. M-government services should be characterized by a "whenever and wherever" usage, that is, the service must be always available and accessed right at the point in time the user requires its functionality.

M-government services should be contextualized and should consider the current situation, or context, of the user during the service utilization. The context of a user, or other entity, is defined by several types of context (sometimes known as context dimensions), such as time, location, activity and preferences (Dey et al., 1999; Schmidt et al., 1998). Context should be explored by m-government services in the sense that it characterizes the user situation and can be used to interpret user's explicit acts, making communication (between users and services) much more efficient.

M-government services should integrate with different external content providers to collect, aggregate and present useful information to users.

3.2 The Report Complaint Scenario

Local governments often have a multitude of disparate internal and external complaint procedures. For example, a pothole in the road may have to be reported directly to the highway maintenance department and then internally passed on to the local road repair team responsible for that section of road. As a citizen it can be rather off-putting to have to locate and

navigate one's way both to and through the appropriate local government offices (or, in some cases, web sites) in order to report such a complaint. It forms a high barrier-to-use. However, in our solution a citizen is expected to be automatically connected to an appropriate complaint service whilst directly at the scene of the trouble, using their own personal wireless device (such as a PDA, mobile phone, or wearable computer).

4. SYSTEM REQUIREMENTS FOR M-GOVERNMENT SERVICES

In this section we analyze the result of transforming the requirements specified in the previous section into a set of system requirements that can affect an appropriate decomposition of the system and the assignment of responsibilities to high-level components. The applied transformational methodology is described in (Machado et al., 2005) and it was applied by a different team and is out of scope of this paper. The derived object model, the USE-ME.GOV platform logic architectural model, is presented in figure 2 . We may identify in figure 2 a Service Repository component (objects {O0a.1.5.d} and {O0a.1.5.i}), a Service Discovery Protocol component (objects {O0a.2.5.c} and {O0a.2.5.i}) and a Context Modeling component (package {P3.}). Such types of components are typically found in service-oriented architectures for the mobile environment (Satyanarayanan, 2001).

4.1 Service Repository and Service Discovery Protocol

Service discovery enables networked entities to discover each other, exchange their functional capabilities, and possibly enter into a relationship. Such component in USE-ME.GOV platform is responsible for discovering, from the known multi-service environment, the appropriate services for accomplish the user request. In order to support efficient service discovery and service aggregation, the service repository component should store the provided services attributes. A proper service description mechanism should be introduced.

We can find in the literature and industrial scene various service-oriented architectures like CORBA (OMG, 2004), UDDI (OASIS, 2003) and Jini (Arnold et al., 1999). These technologies take different approaches in terms of how they support the service discovery process, which result from the various application domains for which they were created or from specific assumptions about their network or computing environments. Particularly, CORBA and UDDI are designed for the Wide Area Network (WAN) setting and Jini for the Local Area Network (LAN) setting. The later is mainly

dependent of particular LAN technologies such as multicast protocols for dynamic discovery of lookup services. In the WAN setting, CORBA objects are potentially disadvantageous because of the constantly changing environment of the Internet. A major restriction of the distributed object-oriented programming approach is that the interactions among objects are fixed through explicitly coded instructions by the application developer. This implies that it is very difficult to reuse an object in a new application without bringing along all its inherent dependencies on other objects (embedded interface definitions and explicit method calls).

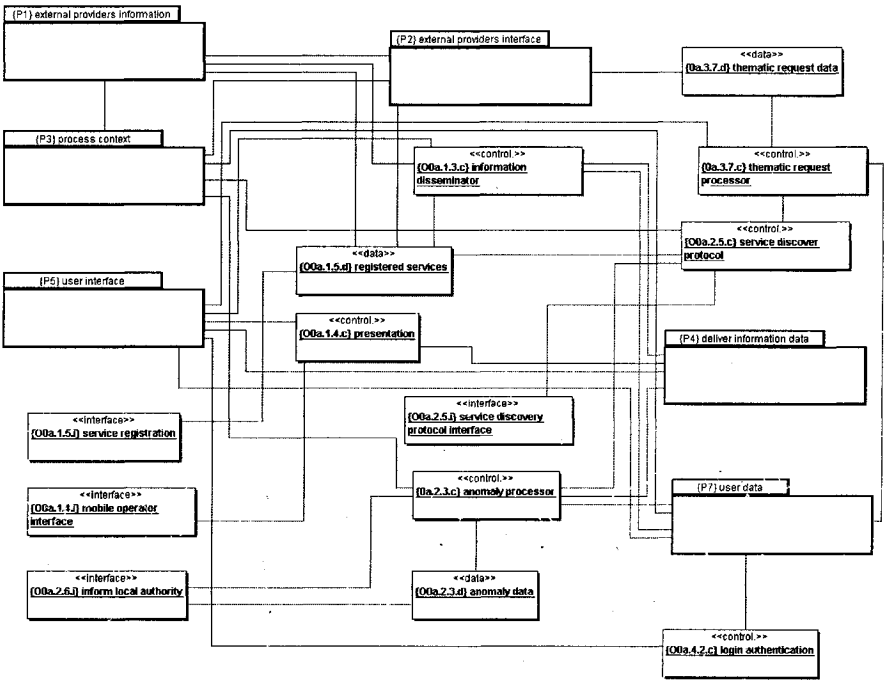


Figure 2. USE-ME.GOV platform logic architectural model

For description purposes, a separate USE-ME.GOV team has designed a meta-protocol of service types based on OWL-S (Bassara et al., 2004). OWL-S is an OWL-based Web service ontology, which supplies Web service providers with a core set of markup language constructs for describing the properties and capabilities of their Web services in unambiguous, computer-interpretable form. OWL-S markup of Web services will facilitate the automation of Web service tasks, including automated Web service discovery, execution, composition and interoperation (World Wide Web consortium, 2004).

4.2 Context Modeling Component

As a result of their mobility, users have a much closer relationship with their surrounding physical environment, and will thus have a much stronger need for information that relates to that environment and their current situation. This need has driven the development of models for supporting the provision of context-aware applications in which information services are associated with particular locations and other context information.

A key research challenge in the development of such systems is how to allow context-aware applications to discover the services they need for the current user situation. Different aspects of context like the position, time, personal preferences or costs should be considered in order to discover a service which matches the user's needs best. Context-aware service discovery is an issue that has been prone to several approaches in recent years, resulting in the development of several standards and technologies. There has been considerable work targeted at supporting location-aware services discovery in mobile computing environments (Cheverst et al., 2000; José et al., 2003; McGrath, 2000), and in context-aware architectures in general (Abowd et al., 1997; Salber et al., 1999). Such systems currently develop individual solutions that only address the requirements of specific application scenarios, are only valid for the specific underlying technology on which they are based or are based on location and context models that are insufficient for associating a service with user situation, mainly because of the different types and scales of known location and context information.

Building context-aware services for mobile environments relies on many different technologies such as context acquisition, context modeling and representation, context reasoning and context-aware service discovery. Mobile information systems maintain many context information providers, and applications or services need to access them to get context information in a homogeneous way. In this way we require homogeneous context service providers' interfaces. Independently developed services and applications must understand context meanings or semantics to advertise themselves to potential users, so we require a formal way to represent context. High-level context information, such as "the user is in a garden", may not be possible to obtain form context service providers, although those provide information that lets the system to infer other contextual information. Mobile users are mainly interested in information that relates to their environment. In this way we require a mechanism that finds the most suitable service to user situation from a wide mobile information service environment.

We are addressing these issues in our work in progress on a context aggregation service (Context Modeling component) (see section 5) that integrates and reasons over objects context information to maintain a

coherent model of objects of the physical and information environment and its location relations, including users, places, coordinates, geographic areas, things and services. Our object model is represented by an object ontology that supports object representation, object classification and its location relationships, as well as context-aware service discovery for context-aware applications.

5. A CONTEXT MODEL FOR SERVICE DISCOVERY

Context information may be derived from elements of the physical world in which we all live, the technical environment in which a software application resides, or even a virtual or conceptual environment that does not, in reality, exist at all. In order to support such a diverse potential range of context information we have designed a flexible object-oriented model that can represent and monitor the parts of those environments that are of relevance to citizens and services. The aforementioned parks, water mains, and graffiti hotspots are examples of real world objects that are represented within our model. Conceptual objects such as complaints and complaint services are also represented within the same model space.

Our model is composed of objects. Everything in the model is an object and anything can be represented by one or more objects. Objects can represent anything from a real-world object to an abstract virtual concept. Objects can also represent the characteristics/properties and relationships of other objects.

An important concept in our model is that any object may be tied to a validity rule: a logical expression that may test and compare the attributes of any object within the model. An object is only visible to external observers if its validity rule evaluates to true, so objects may fade into and out of existence as the model changes over time. Additionally, as the model understands the identity of the observer (in terms of an object within the model) validity rules may be constructed such that certain objects will be valid, and therefore visible, to some observers whilst remaining invalid, and therefore invisible, to others. These observer-dependent validity rules form the backbone of our context-aware service discovery mechanism. For example, a parks complaint service may have a validity rule such that the observer must be within the park that the service is related to. In such a way an observer of the model will only ever see (discover) services relevant for their current context.

A novel approach is also being taken with regard to the representation of location: we view it purely as a relationship between two objects. We call

this relationship the "IsIn" relationship. The "IsIn" relationship is represented as an object, like any other, with attributes that link to the objects representing the entity and the place it is located (not that any object can serve as a place). Chains of "IsIn" relationships can be built up and will typically terminate in the Earth object. The "IsIn" relationship may be parameterized with values to define its precise nature. For example, an "IsIn" relationship between a person and the Earth can be parameterized with a latitude and longitude in order to specify exactly where on Earth the person is. Manipulating different types of location data has long been established as a difficult problem (Leonhardt and Magee, 1996) but our approach offers a flexible solution to combine, compare and reason with location data from different sources and of widely different types.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have argued that citizens should be able to access local government services directly in the time and place in which they are needed. We have presented a set of user requirements for m-government services and the derived logic architectural model for an open service-oriented architecture. We argue that a context modeling component is crucial in supporting context-aware m-government services. We have presented our work on the design of a context modeling component and its application for a report complaint service scenario. As future work several m-government services are being developed - the citizen complaint service is our particular focus - and will be trialed in collaborating municipalities within Portugal, Italy and Poland during 2005.

ACKNOWLEDGMENTS

This work has been funded by the IST project USE-ME.GOV (IST-2002-002294). We would like to thank Ricardo J. Machado, João M. Fernandes e Paula Monteiro from Universidade do Minho their contribution on the formal derivation of USE-ME.GOV platform requirements.

References

Abowd, G., Atkeson, C., Hong, J., Long, S., Kooper, R., and Pinkerton, M., 1997, Cyberguide: a mobile context-aware tour guide, *Wireless Networks*. **3(5)**:421-433.

Arnold, K., O'Sullivan, B., Scheifler, R., Waldo, J., and Wollrath, A., 1999, *The Jini specification*, Addison-Wesley, Reading, USA.

Bassara, A., Filipowska, A., Wisniewski, M., and Zebrowski, P., 2004, USE-ME.GOV project deliverable D5.2.4: Meta-protocol of service types, Poznan University.

CENTURI21 consortium, 1999, The centuri21 project web site (Fev 28, 2004); http://www.centuri21.org.

Cheverst, K., Davies, N., Mitchell, K., and Friday, A, 2000, Experiences of developing and deploying a context-aware tourist guide: The GUIDE project, in: *MobiCom 2000*, ACM Press, Boston, pp 20-31.

Dey, A., Salber, D., Abowd, G., and Futakawa, M, 1999, The conference assistant: Combining context-awareness with wearable computing, in: *3rd International Symposium on Wearable Computers (ISWC'9)*, San Francisco, pp 21-28.

Government of Malta, 2003, M-government web site (May 15, 2005); http://www.mobile.gov.mt.

José, R., Pinto, H., Meneses, F., Vilas Boas, N., Rodrigues H., and Moreira, A., 2003, System support for integrated ubiquitous computing environments, in: *System Support for Ubiquitous Computing Workshop at Ubicomp 2003*, Seattle.

Leonhardt, U., and Magee, J., 1996, Towards a general location service for mobile environments, in: *Proceedings of the 3rd IEEE Workshop on Services in Distributed and Networked Environments*, Macau, pp 43-50.

Machado, R. J., Fernandes, J. M., Monteiro, P., and Rodrigues H., 2005, Transformation of UML models for service-oriented software architectures, in: *ECBS 2005*, IEEE Computer Society Press, Greenbelt, pp 73-82.

McGrath, R., 2000, Discovery and its discontents: discovery protocols for ubiquitous computing, Technical Report UIUCDCS-R-2000-2145, Department of Computer Science, University of Illinois at Urbana-Champaign, April.

Mobhaile project, 2004, The Mobhaile project web site (May 15, 2005); http://www.mobhaile.ie/.

OASIS, 2003, Universal Description, Discovery and Integration of business for the Web (May 15, 2005); http://www.uddi.org/.

OMG, 2004, Common Object Request Broker Architecture (CORBA/IIOP) specification (May 15, 2005); http://www.omg.org/technology/documents/formal/corba_iiop.htm.

Pandora consortium, 2001, The Pandora project web site (May 15, 2005); http://www.ist-pandora.org.

Salber, D., Dey, A., and Abowd, G. D., 1999, The Context Toolkit: Aiding the development of context-enabled applications, in: *CHI'99*, ACM Press, Pittsburgh, pp 15-20.

Satyanarayanan, M., 2001, Pervasive computing: Vision and challenges, *IEEE Personal Communications*, **8(4)**:10-17.

Schilit, B. N., Adams, N. I., and Want R., 1994, Context-aware computing applications, in: *Workshop on Mobile Computing Systems and Applications*, IEEE Computer Society, Santa Cruz, CA, pp 85-90.

Schmidt, A., Beigl, M., and Gellersen, H., 1998, There is more to context than location, in: *Interactive Applications of Mobile Computing*, Rostock, Germany.

USE-ME.GOV consortium, 2002, The USE-ME.GOV project (IST-2002-002294) web site (May 15, 2005); http://www.usemegov.org/.

Visual Admin consortium, 2000, The Visual Admin project (IST-2000-28248) web site (Feb 28, 2004); http://www.visual-admin.net/.

World Wide Web consortium, 2004, Owl-s 1.1 - web ontology language semantics (May 15, 2005); http://www.daml.org/services/owl-s/1.1/.

A METHODOLOGY FOR DESIGNING AND MANAGING CONTEXT-AWARE WORKFLOWS

S. Modafferi[1], B. Benatallah[2], F. Casati[3], and B. Pernici[1]

[1] *Politecnico di Milano*
Dipartimento di Elettronica e Informazione
P.zza L. Da Vinci 32 - 20133 - Milano Italy

[2] *University of New South Wales,*
CSE, Sydney NSW 2052, Australia

[3] *Hewlett-Packard, Palo Alto, CA, 94304, USA*

Abstract: The increased availability of context information and the widespread adoption of more and more powerful devices creates the opportunity and desire for context-aware applications. In this paper we focus on a specific but important type of applications: workflow applications. Just like other applications, workflows too require context-aware capabilities, that is, require the capability of modeling business logic that is sensitive and varies depending on the users' context. In this paper, we propose a methodology for context-sensitive business processes development. We extend existing process modeling languages to allow modelling context sensitive regions (i.e, parts of the business process that may have different behaviours depending on context). We also introduce *context change patterns* as a mean to identify the contextual situations (and in particular context change situations) that may have an impact on the behaviour of a business process. Finally, we propose a set of transformation rules that allow generating a BPEL-based business process from a context sensitive business process. This allows using existing process engines to support context-sensitive business processes.

Keywords: User Context, Context-aware Workflow, Adaptive Workflow, Context Sensitive Regions, Dynamic Workflow Execution

1. INTRODUCTION

Capturing and managing user context is becoming more and more important in business applications. With recent advances in mobile technologies that provide several ways for identifying user contexts (e.g., identifying user location) (Capra et al., 2003; Chakraborty and Lei, 2004;

Dey and Abowd, 2001) and for leveraging context-based applications, the need and opportunity for delivering customized services to users in different situations is becoming prominent (Roman et al., 2000).

Generally speaking, *context* refers to information that characterizes the situation of a person, place, or object, that is relevant to a given system. In our work, we focus on context information that is relevant to provide personalized services to users based on their environment and needs. This includes information such as user location, current user device, or network bandwidth. Existing work in delivering personalized services focused mostly on capturing and representing context information (Chakraborty and Lei, 2004; Dey, 2001). This is clearly an interesting and necessary step as it provides for recognizing context as a separate abstraction, and fosters the development of tools and techniques for context management.

However, delivering personalized services to users requires methodologies and techniques that not only allow developers to capture and manage context, but that also facilitate the creation of context-aware applications. As a particular and very important kind of (context-aware) applications, in this paper we consider *workflow* applications, and more in general applications developed using composition technologies (also called process technology in the following). Workflow applications are rapidly gaining popularity, especially with the advent of Web services and the push towards service-oriented computing, as the availability of homogeneous components (services) makes it easier to develop applications by composing existing building blocks.

Specifically, the aim of this paper is to identify techniques that facilitate the development of context-sensitive processes, including in particular the ability to manage context change, a key issue in any context-aware application development. These techniques are based on what we believe to be simple but essential extensions to "traditional" process models. Indeed, the main challenge in this work has been that of identifying the process modeling concepts that could capture the essence of context-sensitive applications, or at least a wide variety of them, while avoiding to unnecessarily making process modeling more complex.

Our work builds upon existing techniques for managing context information and extends them by providing a methodology and an architecture for developing context sensitive processes. We propose the concept of *context-sensitive region* to localize parts of the process that have different behaviors according to context. A set of *context change patterns*, that classify and capture typical context changes is defined. Context-sensitive regions include the description of how to react to a context change defined in a context change pattern. The proposed model will

be used as a basis to generate the BPEL process and exception handlers to manage context-sensitive processes.

The paper is structured as follows: in Section 2, we discuss the integration of context in business process models. Section 3 presents our approach for modeling context-sensitive business processes. The runtime aspects are presented in Section 4. Section 5 presents related work, while Section 6 gives conclusions and future directions.

2. CONTEXT-SENSITIVE BUSINESS PROCESS MODELING

2.1 Context Definition

Delivering personalized services requires a precise definition of user needs and user environments. Traditional service provisioning relies on a relatively static characterization of the user and the user's context, because the changes in the user environment are relatively limited and because of the inability to dynamically and automatically capture context and context change (Roman et al., 2000).

Several definitions of context have been proposed in the literature. In (Dey, 2001), context and context-aware computing have been defined as:

> *"Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves."*
>
> *"A system is context-aware if it uses context to provide relevant information and/or services, where relevancy depends on the user's task."*

In general, the user context can consist of many different aspects, and different context models include different properties as elements of a context. The most common and typical context attribute is the geographical location, which can be expressed at different levels of granularity (XYZ coordinates, city, state, etc). Other context information may include a "logical" location (e.g., "in a meeting" or "at home"), the present occupation, the weather at the user's location, and many other attributes.

For the specifics of the context model, the present paper is based on the one developed within the MAIS project (Cappiello et al., 2005). In particular, besides generic location-related user attributes, the MAIS model focuses on user access devices (e.g., PDAs, laptop, or others) and quality of network connectivity. However, it should be noted that while we consider a specific context model as example, the concepts presented in the paper are generally applicable to virtually any context model. In fact, even if in our approach workflow design is influenced by consideration of context, the use of high level representation of context

decouples the problem of managing workflow schema modification and of context model evolution.

2.2 Context-Sensitive Process Models

There are many ways in which processes can leverage context information to create context-sensitive applications. One is to make routing decisions. For example, information about skiing conditions can be delivered to users that are in a mountain area, but not to others. Another usage of context information is to configure a given service invoked as part of the process, or to select a specific provider among the ones able to provide a given service. For example, a news delivery service may be configured to send videos of different quality based on the network bandwidth or user screen resolution. Finally, an important aspect is context change management, which involves the ability of the process to alter its course based on new context information about the user to which the process is delivering a service.

In prior work, there is a strong relationship between application and context model and very often they are developed together.

In our approach, the user context is considered as first class citizen in business processes. Each process has an implicitly defined set of data elements (process variables) that capture the user context. These variables are automatically populated and maintained by the infrastructure. For example, the user context may include the notion of "user device", which can assume the values of PDA and PC. The context is monitored by a *context monitor*, which measures context variables according to what is called *Low level context model* (see Fig. 1). We use the term "Low level" to denote that it is typically at a low level of abstraction (e.g., a location can be expressed in xyz coordinates). For this a conventional context representation model is used (i.e., (Cappiello et al., 2005; Chakraborty and Lei, 2004)). This model provides means to represent user context attributes and user context changes (also called context change events), such as decrease in throughput or change in location. The low level context is then mapped to high level context *dimensions*, which form the high level user context (or simply user context hereafter). This mapping is necessary to abstract from the details of the context (and from context changes) that would be too finegrained and detailed for most workflow applications. Each dimension can have a set of predefined values (High, Low; PDA, PC), and correspond to the process variables that are implicitly included in each workflow models.

The Context Mapping module is in charge of translating measured context values into values for the dimensions at the higher level, accord-

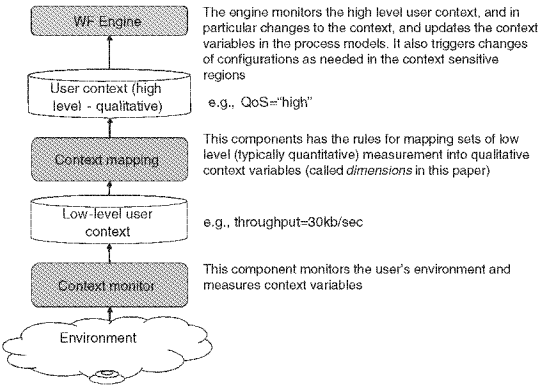


Figure 1. Relationship between workflow and context

ing to a set of Context Rules. For instance, if in the low level context the variable representing the throughput of the net decreases, this fact is transparent to the High Level User Model until a given threshold is reached, after which a rule in the Context Mapping component determines that the throughput dimension has now the value"Low Qos".

To facilitate context-sensitive process modeling we also propose in the workflow model the concept of *context-sensitive* region. A context-sensitive region is a part of the business process that may have different behaviors (e.g., different flow structures) depending on the context. A region is associated to several configurations - essentially corresponding to subprocesses - that represent the different behaviors. Whenever the region is instantiated, a specific configuration, among the possible ones, is selected based on the context.

The designer has to define an "entry condition" for each region. It is a boolean condition defined over the user context. The set of entry conditions of configurations for a region must define a partition over the user context. An"otherwise" condition can also be specified to capture all cases in which no entry condition is verified. Note that if the entry conditions do not define a partition, then the behavior of the system is non deterministic (the system currently does not enforce that conditions create a partition of the user context).

The problem of capturing and managing context change is at the heart of context-sensitive applications. The key here is to devise a model that makes it easy for developer to manage simple context changes, possibly with minimal or no modeling, while giving also the possibility to manage

complex context change requirements. To this end, we propose an approach that is based on schema evolution techniques and migration rules. Schema evolution is the default context change behavior provided by the model, and the nice aspect of it is that it requires no modeling from the user. Briefly, it works as follows: as mentioned, the system takes care of automatically monitoring the context. This also happens when a workflow instance is within a CSR. At some point, it may be that the system detects that the entry condition for the CSR in which an instance is (called "initial configuration") turns out to become false, and the entry condition of another configuration (called "target configuration") becomes instead true (one entry condition must be true as conditions identify a partition). In this case, the problem is handled analogously to what is done in workflow schema evolution: the execution of the instance is rolled back until a point in the CSR is reached where the execution of the instance in the initial configuration "can be seen as" an execution of the instance in the target configuration, at which point the instance is migrated to the target configuration and continues as prescribed by the process flow in the target configuration. As a simple example, the initial configuration can be composed by a sequence of tasks A,B,C,D, while the target configuration is a sequence of tasks A,B,X,Y. If the context change occurs while the instance was performing task D, then the execution of D is aborted, and the execution of C is compensated. At this point, the execution of the instance in the initial configuration (A,B) can be seen as an instance of the target configuration (which also starts with A,B). Hence the instance is migrated and the next task to be activated is X, according to the flow in the target configuration. In the worst case, the compensation of the initial configuration continues to the start of the CSR. We do not detail schema evolution further as it has been the subject of many papers (although in different scenario - our contribution here is its application to context change management and the idea to use this approach to simplify process definition and maintenance). The interested reader is referred to (Casati et al., 1998; Reichert et al., 2003).

Schema evolution proposes a default behavior that may be inadequate in some situations, for example due to the excessive loss of work due to compensation of completed activities. To handle this cases, our approach proposes the notions of *context change patterns* and of *migration rules* to respectively capture the interesting change events and manage the change with ad hoc behavior Context change patterns capture the different ways in which a context can evolve and help defining how the change is relevant from a process model perspective. Patterns help characterizing and classifying the different types of context changes and as such simplify the definition of how to handle a change within a

context-sensitive region. Indeed, the combination of context regions and context change patterns allows for flexible and easily maintainable process models, which are even robust to the evolution of the context model itself. A Transition is an instantiation of a Pattern. It causes changes within active workflow instances, and specifically in the context-sensitive regions sensible to the related context-change pattern. For a given transition in the User Context, the user may associate a Transformation in the Workflow model. A transition is analogous to an ad hoc migration. It includes the explicit identifications of points in the workflow where an instance should migrate from the initial to the final configuration (regardless of whether at that point the execution in the initial configuration can be seen as an instance of the target configuration), along with rules to manage the migration (e.g., data transformation).

2.3 Context Change Patterns

To enhance the design of context-aware workflows, it is important to determine patterns that capture common behaviors followed by the users to which the workflow provides a service. We identified four patterns which cover several interesting situations as well as a framework for building new patterns. As stated in the previous paragraph, patterns in our model are the basis for context state transitions (e.g. for the pattern "Device dependency" the transition can be "PDA to Pc" and "PC to PDA"). It is possible to define new interesting dimensions and then new pattern as transitions in this dimension. The identified patterns are:

Device dependency. It represents the situation when a user changes device, but is still involved in a business process. For example, the user uses a PDA initially, being out of office, and then connects through a desktop later on, when in the office, continuing to interact with the same process. The specificity of this pattern is that it could provide for data movement; in fact during a process some information can be stored only on the client side and, changing the client device, it is necessary to perform a migration of this data to the new device.

Qos driven choice. It represents the possibility of having different business process configurations according to different QoS levels in the client side. A typical situation is the abrupt decrease of available bandwidth that requires a substitution/reconfiguration of some services on the server side.

Location driven choice. It represents a situation where business process configurations are associated to given user locations. For example, the selection of a restaurant from a list will be different depending on the current city and country, offering different selection services to

the user and, if the location based choice is enough grained, a user can change location within a business process several times.

User on-line/off-line. Due to resource consumption constraints, user may prefer to work off-line. This pattern represents the typical working mode of a user connected through a mobile device. The interaction of wfms engine with the user is not continuous and during the user absence the process can show a different behavior waiting for his return.

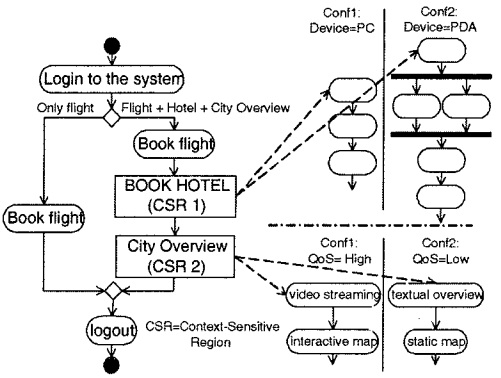


Figure 2. Wf with regions

2.4 Example

We illustrate the proposed approach using a simple example, taken from the travel reservation domain. In the sample process, a user can login in the system, and is then faced with the following two possibilities: i) book a flight; ii) book a flight, a hotel, and also get information about the destination city.

Assume that the flight booking task is context independent while the other tasks are context-sensitive (i.e., they have different behaviors according to the current user context). Fig. 2 shows the model of such business process.

The "Book Hotel" Region (CSR1) is sensible only to the Pattern *Device Switching* and thus the interesting Dimension is only "Device"; in the user context model there are two states "PC" and "PDA" and in the workflow model two configurations for the region. The difference between configurations are mainly that if the user device is a PDA the server will send in parallel part of the information to the PDA and part to a different client: a mailbox.

The "City Overview" Region (CSR2) is sensible only to the Pattern *Qos driven choice* and thus the interesting Dimension is only "QoS"; in the User Context Model there are two states "High QoS" and "Low QoS" and in the Workflow Model two configurations for the Region. The difference between configurations are that if the QoS is high the services are interactive (e.g. dynamic map), otherwise they are simple static information (e.g. textual information). Further details about the example will be presented in Section 3.2.

3. MODELING ABSTRACTIONS

In this section we will introduce our model. The Workflow model is composed of Context Sensitive Regions and each region is composed by different configurations. Also context model aspects are discussed.

3.1 Workflow Model

To describe the workflow model we define a workflow as directed graph. Our definition is based on traditional definition of workflows as graphs and then only specific constructs are presented here. In our model a Workflow W is composed by a set of i) Traditional coordination C and task T nodes; ii) Regions Reg; iii) Directed flow arcs F_W.

Let $Wgraph =< NW, FW >$ be a graph where NW: finite set of nodes, FW: directed arc (flow relation) $FW \subseteq NW \times NW$

$\forall nw \in NW, Nodetype : nw \rightarrow \{Coordinator, Task, Reg\}$

$NW = C \cup T \cup Reg, \quad (C \cap T = \emptyset) \wedge (C \cap Reg = \emptyset) \wedge (T \cap Reg = \emptyset)$

where C: set of coordinator nodes (e.g. switch, loop), T: set of Task nodes, Reg: set of Context Sensitive Region.

Context Sensitive Region. A Context-sensitive region (hereafter CSR) is a subprocess of the workflow that may have several configurations exporting different behavior according to specific conditions (i.e. user context).

A region is composed of alternative configurations linked with particular arcs called *migration arcs.* A migration arc is associated with instructions on how to migrate a workflow instance from one configuration to another.

A typical example of information associated to a migration arc is the data items that need to be moved when a context change occurs.

For example the *Device Switching* pattern provide for the change of the device and this fact can imply the necessity to move some data from the old device to the new one (see also the example in Section 3.2).

Configuration. A Configuration is a subprocess of the workflow related to a set of states in the User Context Model (i.e. $User\ Device = PDA$). A Configuration $Conf_i$ is composed by: i) an entry condition EC; ii) normal coordination C and task T nodes; iii) a set of Starting Migration Points MPs; iv) a set of Ending Migration Points MPe; v) a set of directed Configuration Arcs FC.

The entry condition EC is an expression used to define when the configuration has to be entered.

To handle the change of configuration there is a default behavior. This default behavior requires no modeling and no effort by the user. The system migrates the workflow from a configuration C1 to another C2. Migrating essentially means rolling back the execution of C1 up to the point where the instance can be seen as an instance of C2, as for the work in Workflow evolution (Casati et al., 1998). At that point the instance is migrated and the process continues. This also means that each task in this model has to be associated to a (possibly null) compensation action.

For the most complex cases, the user can explicitly specify migration rules. These rules are modeled using migration points and migration arcs. A starting migration point is a point in a configuration $C1$ where a direct link, that is a migration arc, to a new configuration is available. Assume the flow is in a configuration C1 and has to migrate to a configuration C2, if in the rolling back process a migration point and a migration arc from C1 to C2 is found then if is followed to perform the transformation, otherwise the default behavior is exploited. The definition of migration point is similar to the notion of safe-point in the WIDE project (Grefen et al., 1999).

Migration points and migration arcs provide high expression power to specify migration rules. For example they allow the designer to define business equivalent behaviors only by connecting two points and, in case, associating a process to the arc. This power has a cost in terms of what the designer has to specify, so we give this possibility in our model, but we expect the default behavior to apply often and we do not realistically expect a CSR to have a large number of configurations. Hence, this model, and in particular the combination of automated (default) migration along with the possibility of specifying migration rules for special cases is able to combine the need for ease of modeling/manageability with the possibility of defining specific migration semantics for the most complex cases. Now a formalization of the configuration is provided.

Let $Conf =< EC, NC, FC >$ be a graph where EC: the expression for the entry condition for the configuration, NC: finite set of nodes, FC: directed arc $FC \subseteq NC \times NC$

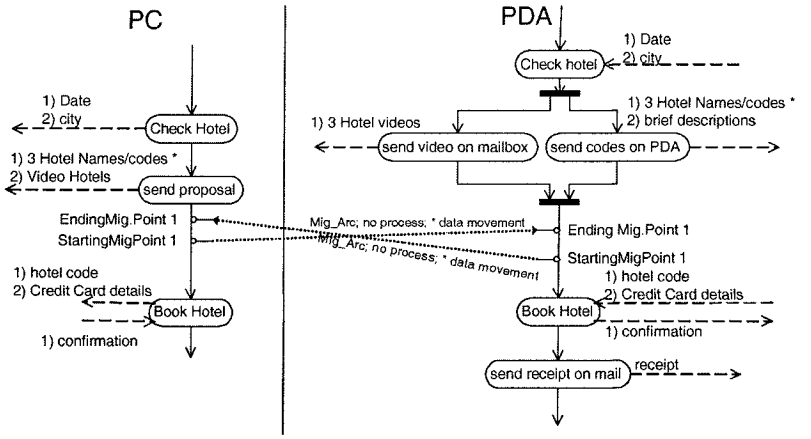


Figure 3. Region Book Hotel configurations for device switching pattern

$\forall nc \in NC, Nodetype : nc \rightarrow \{Coordinator, Task, StartingMigration Points, EndingMigrationPoints\}$

$NC = C \cup T \cup MPs \cup MPe,$
$(C \cap T = \emptyset) \wedge (C \cap MPs = \emptyset) \wedge (C \cap MPe = \emptyset) \wedge (T \cap MPs = \emptyset) \wedge (T \cap MPe = \emptyset) \wedge (MPs \cap MPe = \emptyset)$

where C: set of coordinator nodes, T: set of Task nodes, MPs: set of Starting Migration Points, MPe: set of Ending Migration Points.

Formalization of CSR. Let $Reg_i =< \bigcup_i Conf_i, MA >$ be a graph where $Conf_i$: a Configuration, MA: directed arc $MA \subseteq MPs \times MPe \times Op \times Proc$ where Op: set of optional Operations, $Proc$: set of optional Processes.

3.2 Example

Let us suppose to design a process starting from scratch. The first step of our methodology is the definition of the high-level workflow already presented in Section 2.4 and shown in Fig. 2. This step provides a high-level summary of the entire workflow and allows considering in the further steps each region as a stand alone part. The second step is the definition of each CSR.

Book Hotel CSR (see Fig. 3). This CSR is sensible to "Device Switching" pattern. It shows two different behavior according to the used device.

If the user is using **PC**, after the reception of a query, the server will send him information about some hotels and also a video. Then it will

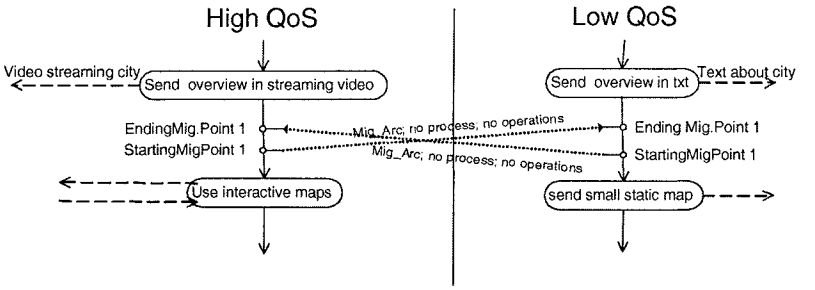


Figure 4. Region City Overview configurations for QoS driven pattern

provide the task for booking according to codes provided by the user. All operations are in sequence.

If the user is using **PDA**, after the reception of a query, the server will send him on the PDA some brief information about hotels and will send on the user email the video of the hotels. These operation are carried out in parallel. Then the server will wait for the hotel code, provided by the user, to make the reservation. Finally it will send a receipt on the user email. There is not a task for receipt in PC configuration because we suppose that the interaction is by web-browser and so the printout of the page is enough as receipt. It is less probable that a PDA is connected to a printer and so we provide a receipt on email box.

In Fig. 3 there are a pair of migration arcs. Whichever is the transformation (PC to PDA or PDA to PC) starting from MPs1 (Starting Migration Point 1) the migration arc is associated with an operation for "data movement". This operation allows moving data on the client side (in this case the hotel codes marked with a * in the figure). In fact, even if, the process is running on the server side, it is possible to have data movement in the client side. The optional process associated with the migration arc represents a process related to the configuration change and it is defined considering the two configurations and the starting, respectively ending, migration point.

The default exception management is used if the user context changes before the flow meets a migration point, that is schema evolution rules are applied. In the other situations the migration is driven from the migration arc starting from the last migration point. In Fig. 3 the migration arcs expresses the business equivalence of operation executed until the migration point. The system can not determine automatically this equivalence, in charge of the business logic, and thus the designer had to explicitly define it.

City Overview CSR (see Fig. 4). This CSR is related to the pattern "QoS driven choice". We assume that the user device configuration is sensible to different level of QoS (i.e. he is using an UMTS connection). With **High QoS** the first task is used to send a video streaming about the city. Then the user can use interactive city maps.

With **Low Qos** the first task is used to send some small texts about the city. Then the server will send to the user a small static city map.

In this CSR all the migration arcs are not associated with processes or other operations. Here the migration arcs are used to express the equivalence in terms of behaviours. Differences are in terms of realizations of the same behaviours according to different contexts.

4. RUN-TIME ARCHITECTURE

The output of the design phase will be a WS-BPEL process annotated with context abstractions to represent context-sensitive regions. . Each region is modeled as a stand-alone part. We propose transformation approach to translate a context sensitive WS-BPEL process into a conventional WS-BPEL process with the appropriate handlers.

We assume that a service oriented architecture is used to manage user context (this kind of approach is followed in several systems, e.g (Chakraborty and Lei, 2004)). In our approach the context manager is designed as a web service. Context changes are conveyed to the WS-BPEL engine as messages (context change event, with parameters) from the context manager. Context changes are managed in the WS-BPEL as exceptions and therefore the unified BPEL is composed of a normal behavior (conditional according to the context) and a context change handler to manage context exception.

The handler checks if an automatic migration (following the traditional schema evolution methods) is available. If it is not available the handler performs a reverse path compensating each task until reaching the last starting migration point, then it performs the actions and processes associated with the migration arch leading to the new configuration and, eventually, it restarts the normal flow from the ending migration point (in the new configuration) reached with this migration.

Fig. 5 shows the Ws-BPEL unified schema (with specific operation for region switching) that is the output of the CSR shown in Fig. 3 after the automatic generation process. This schema is referred to a standard WS-BPEL and the corresponding behavior is context-dependant according to the design presented in the previous sections.

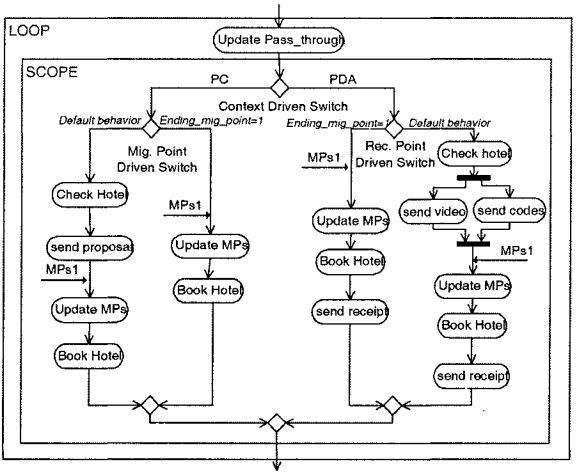


Figure 5. Context Sensitive Region of Fig. 3 as unique WS-BPEL schema

5. RELATED WORK

The possibility of modifying workflows to increase their flexibility is widely studied (Casati and Shan, 2001; Müller et al., 2004; Reichert et al., 2003; Shan et al., 2005). The reason of workflow modification are various and not strictly related to the user "context" issue. The context-aware workflow issue concerns many and different aspects that are now being linked together. The most common definition of context is provided in (Dey, 2001); generic context models like (Chakraborty and Lei, 2004; Dey and Abowd, 2001) can be the basis of context-aware applications; different middleware have been proposed to interpret the context providing useful information to the applications (Bellavista et al., 2003; Capra et al., 2003). A way followed to build context aware applications can be the definition of self-contained systems like (Zariskas et al., 2001). They have their own definition of context, are mobile-oriented and context-aware, but they do not focus workflow systems. Other systems are workflow-based, but they focus on single context-sensitive tasks (Long et al., 2004; Patil et al., 2004; Sheng et al., 2004).

The work presented in (Binemann-Zdanowicz et al., 2004) proposes an approach to context-awareness for Web Information Systems that distinguishes among the various kinds of contexts, but it is not clear how it manages together workflow execution and its rich context model.

Some recent papers propose the extension of workflow languages and models with aspect-oriented software design (Charfi and Mezini, 2004). They suggest the usage of aspect-orientation as a complementary tech-

nique for workflow modelling and specification presenting a hybrid approach for realizing the integration of business rules (modelled as aspects) with a WS-BPEL orchestration engine by using aspect-oriented programming techniques. Our definition of high level context can be viewed as an Aspect and the proposed solution is how to exploit these different aspects by using a standard, even if annotated, workflow model and a standard workflow engine.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a methodology for designing and developing Context-sensitive business processes.

By identifying patterns for capturing context changes, our approach provides a loose coupling between workflow definition and context model. By means of Context-Sensitive Region, a high level construct for modelling context changes is provided and, eventually, by annotating an existing workflow model a solution based on the currently available WS-BPEL for the run-time management is provided.

We are studying a more general use of regions defining region-level properties that are specific to a context, but not necessarily related to the fact that a context may change. In fact even if the context is always the same from start to end in a given workflow execution, it is possible to define region-level properties that depends on the context, e.g., attributes such as data transfer rates or resolution, used by all activities in the region.

ACKNOWLEDGEMENTS

This work is partially funded by the Italian MURST-FIRB MAIS Project (Multi-channel Adaptive Information Systems) and by a visiting research grant from School of Computer Science and Engineering of the University of New South Wales (UNSW) Sydney.

References

Bellavista, P., Corradi, A., Montanari, R., and Stefanelli, C. (2003). Dynamic binding in mobile applications: A middleware approach. *IEEE Int. Computing*, 7(2):34–42.

Binemann-Zdanowicz, A., Kaschek, R., Schewe, K., and Thalheim, B. (2004). Context-aware web information systems. In *Proc. of Asia-Pacific Conference on Conceptual Modelling*, pages 37–48, Dunedin, New Zealand. Australian Computer Society.

Cappiello, C., Comuzzi, M., Mussi, E., and Pernici, B. (2005). Context management for adaptive information systems. In *In Proc. of Int. Workshop on Context for Web Services (CWS)*, Paris, France.

Capra, L., Emmerich, W., and Mascolo, C. (2003). Carisma: Context-aware reflective middleware system for mobile applications. *IEEE Transactions on Software Engineering*, 29(10):929–944.

Casati, F., Ceri, S., Pernici, B., and Pozzi, G. (1998). Workflow evolution. *Data Knowl. Eng.*, 24(3):211–238.

Casati, F. and Shan, M. (2001). Dynamic and adaptive composition of e-services. *Information System*, 26(3):143–163.

Chakraborty, D. and Lei, H. (2004). Pervasive enablement of business processes. In *Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications (PERCOM)*, Orlando, (Fl), USA.

Charfi, A. and Mezini, M. (2004). Aspect-oriented web service composition with AO4BPEL. In *Proc. of European Conference on web Service, (ECOWS)*, pages 168–182, Erfurt, Germany. Springer.

Dey, A. (2001). Understanding and using context. *Personal Ubiquitous Computing*, 5(1):4–7.

Dey, A. and Abowd, G. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-Computer Interaction Journal*, 26(2-4):97–166.

Grefen, P., Pernici, B., and (Eds), G. S. (1999). *Database Support for Workflow Management - The WIDE Project.* Kluwer Academic Publishers.

Long, Y., Lam, H., and Su, S. (2004). Adaptive grid service flow management: Framework and model. In *Proc. of IEEE Int. Conf. Web Services (ICWS)*, pages 558–565, San Diego, Ca, USA.

Müller, R., Greiner, U., and Rahm, E. (2004). AGENTWORK: A workflow-system supporting rule-based workflow adaptation. *Data and Knowledge Engineering.*

Patil, A., Oundhakar, S., Sheth, A., and Verma, K. (2004). Meteor-S web-service annotation framework. In *Proc. of Int. Conf. WWW*, pages 553–562, New York, NY, USA.

Reichert, M., Rinderle, S., and Dadam, P. (2003). ADEPT workflow management system. In *Proc. of Int. Conf on Business Process Management BPM*, pages 370–379, Eindhoven, The Netherlands. Springer.

Roman, G., Picco, G., and Murphy, A. (2000). Software engineering for mobility: a roadmap. In *In proc. of Inter. Conf. on Software Engineering (ICSE) - Future of SE Track*, pages 241–258, Limerick, Ireland.

Shan, E., Casati, F., and Dayal, U. (2005). Adaptive process management. *to appear in Int. Journal on Business Process Management.*

Sheng, Q., Benatallah, B., Maamar, Z., Dumas, M., and Ngu, A. (2004). Enabling Personalized Composition and Adaptive Provisioning of Web Services. In *Proc. of Int. Conf. on Advanced Information Systems Engineering (CAiSE)*, pages 322–337, Riga, Latvia. Springer.

Zariskas, V., Papatzanis, G., and Stephanidis, C. (2001). An architecture for a self-adapting information system for tourists. In *Proc. of the Workshop on Multiple User Interfaces over the Internet: Engineering and Applications Trends (in conjunction with HCI-IHM')*, Lille, France.

AN EXTENSIBLE TECHNIQUE FOR CONTENT ADAPTATION IN WEB-BASED INFORMATION SYSTEMS

Roberto De Virgilio and Riccardo Torlone
Dipartimento di Informatica e Automazione
Università Roma Tre
Roma, Italy
{devirgilio,torlone}@dia.uniroma3.it

Abstract: Adaptive Web information systems need to exploit information about the context of the client in order to deliver, in an appropriate way, relevant information. In this paper, we present a general approach to this problem that is able to handle heterogeneous context information and different coordinates of adaptation. The approach is based on a general notion of profile that can be used to represent a variety of contexts at different level of details. Client profiles, possibly expressed in different formats, are wrapped and translated into such a general representation. The analysis of profiles drives the generation of an interface configuration. A configuration specifies, at the various layers of a Web based Information System (content, navigation and presentation), how to build a response that meets the requirements of adaptation of the profile. We describe architecture and functionality of a prototype implementing such adaptation methodology and illustrate practical examples of use of the prototype.

Keywords: Adaptive Information Systems, Context awareness, Data Intensive Web Sites, Heterogeneous profiles

1. INTRODUCTION

An ever increasing number of mobile devices, such as PDAs and next generation phones, can be deployed today to provide *everywhere* and *any time* access to the Web. This scenario presents a number of challenges, especially when it comes to data-intensive applications, mainly because these devices offer limited computing capabilities. It follows that a novel and fundamental requirement of modern Web based Information Systems is the ability to adapt and personalize content delivery according to the *context* of the client (a human being or an application). There

is no accepted definition of context, but the term is a usually adopted to indicate "a set of attributes that characterizes the capabilities of the access mechanism, the preferences of the user and other aspects of the context into which a Web page is to be delivered" (W3C Device Independence Working Group, 2003). These may include the access device, the network QoS, the user preferences, the location, and so on.

Many approaches have been proposed to the problem of the design and the implementation of an adaptive Web based information system (Bickmore et al, 1999; Ceri et al, 2000; Ceri et al, 2003; Kießling, 2002; Leubner and Kie, 2002; Gu and Helal, 2004; Pastor et al, 2003; Schwabe et al, 1996). However, they are often specific solutions, suited only for predefined adaptation requirements (usually device characteristics and user preferences) and hardly reusable for adding new adaptation functionality to existing systems. In these approaches a relevant problem that has received little attention is the large heterogeneity of formats used to express context information: text files in ad-hoc format, HTTP headers, XML files over specific DTD's, RDF (Lassila and Swick, 1999; Brickley and Guha, 2003), CC/PP (Klyne et al, 2004).

In our study, we focus our attention on the large category of *data intensive* Web Information Systems, which mainly provide a Web access to large amounts of structured data, and address the problem of providing a general solution to the problem of content adaptation that can be used for different and possibly independent requirements of adaptation.

To this end, we first present a very general notion of *profile* that can be used to represent a large variety of aspects of a context, at different level of details. Each profile is associated with a *configuration* that specifies, in an abstract way, how the Web pages that compose the response to deliver to the user have to be generated. This is done by taking into account both the specific user request and his/her context. Profiles can be compared and configurations can be merged: these features are used for reusing configurations specified for "compatible" profiles and for integrating different requirements of adaptation.

The activities of the various components of the architecture are coordinated by a general methodology for content adaptation, based on the generation and management of configurations. In order to experiment the effectiveness of our approach, we have designed and developed a prototype implementing the proposed methodology. The tool is based on a very general architecture for an adaptive system that can be easily extended to meet new requirements of adaptation, not fixed a-priori. The tool makes use of a caching technique that take advantage of a repository of configuration implementations to make the adaptation process more efficient. A involved component of the tool is the profile interpret

that takes as input incoming profiles expressed in a variety of formats and converts them in a common format that represents the context in terms of our general model of profile.

The paper is structured as follows. In Section 2 we present a general model for adaptive Web-based information systems. In Section 3, we present the general methodology of adaptation, by illustrating the basic concepts of *profiles* and *configuration* and showing their use in a practical example. In Section 4 we first present a general architecture and then illustrate a practical implementation of the system based on this architecture. Finally, in Section 5 we draw some conclusions.

2. A MODEL FOR ADAPTIVE IS

In this section we present a simple but general model that we use as a reference for our approach. It is based on two main "logical" constructs: the profile and the configuration.

2.1 Profiles

A *profile* is a description of an autonomous aspect of the context in which the Web site is accessed (and that should influence the presentation of its contents). Examples of profiles are the user, the device, the location, and so on. A *dimension* is property that characterizes a profile. Each dimension can be conveniently described by means of a set of attributes that associates actual values with a dimension. For instance, a profile for a client device can be represented by means of the hardware, software, and browser dimensions. In turn, the hardware dimension can be described by means of attributes like CPU, memory, and display. Within a dimension, it is useful to introduce a partial order between attributes at different level of details. As an example, in Figure 1 are reported possible dimensions for the profile of a device. In the HARDWARE dimension, the HD spec attribute fully describes the hardware capabilities of the device, whereas the Display attribute just represents the display dimensions of the device. Possible profiles over these dimensions are reported in the bottom of the same figure.

A *context* is simply a collection of profiles. Note that this notion of context is very general and is therefore suited to model almost all context formalisms proposed in the literature and adopted in practical systems.

In our model, different profiles over the same dimensions can be compared making use of a subsumption relationship $\lhd$. Intuitively, this relationships compares the level of detail of the profiles: given two profiles P_1 and P_2, we say that P_1 subsumes P_2 ($P_1 \lhd P_2$) if P_2 includes all the attributes of P_1 at the same or at a coarser level of detail, according to

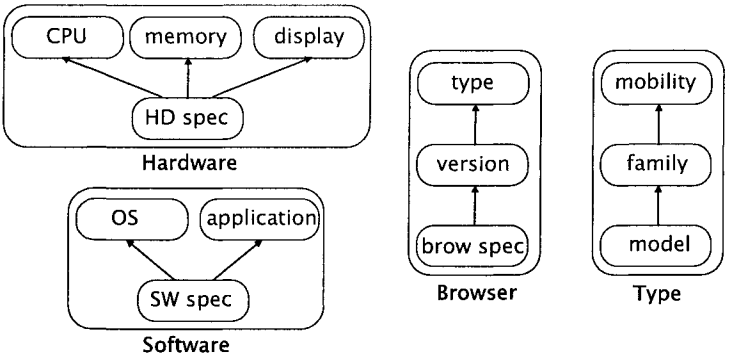


Figure 1. Dimensions for a device and possible profiles over them

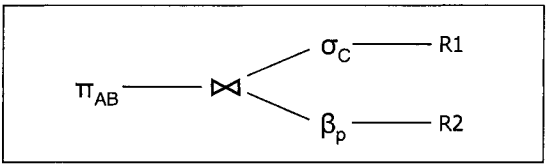
the hierarchy defined on the attributes of the dimensions. We assume that $\lhd$ has a unique top element called the *generic* profile. As an example, the profile DeviceType in Figure 1 subsumes the profile DeviceRed that, in turn, subsumes the profile DeviceFull.

2.2 Configurations

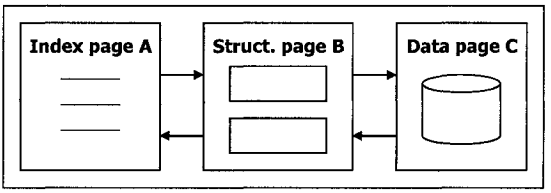
Let us now turn our attention to an adaptive Web-based Information System (WIS) that is able to modify and personalize delivery of contents and services according to different profiles. As we have said in the introduction, we focus our attention on the large category of data intensive WIS, which mainly provide a Web access to large amounts of structured data. In such applications, it is useful to consider separately its three main components: the content (that is, the data to publish), the presentation (that is, the layout of the pages) and the navigation (that is, the hypertext structure of the Web site). It is agreed (Fiala et al, 2003; Fiala et al, 2004; Frasincar et al, 2004; Vdovjak et al, 2003) that an adaptation process should operate on all these components: selecting the most appropriate content (e.g., according to user interests), building an adequate layout for the web pages (e.g., according to the layout capabilities of the client device) and organizing the hypertext structure of the web interface (e.g., decomposing large contents in linked pages, when the band of the communication channel is limited). This approach has been

According to this observation, we introduce the abstract notion of *(interface) configuration* as a triple $C = (q, h, p)$ where:

q



h



p

Text	Link	Image	Video
Font: Arial	**Note:** False	**Format:** jpeg	**Format:** mpeg
Size: 10pt	**Font:** Arial	**Size:** 178 X 208	**Size:** 240 X 320
Style: Normal	**Size:** 10pt	**Border:** 0pt	**Border:** 2pt
Color: Black	**Style:** Underline	**Color:** True	**Color:** False
Border: 0pt	**Color:** Blue	**Alignment:** left	**Alignment:** center

Figure 2. A configuration $C = (q, h, p)$

- q is a query over the underlying database expressed in standard relational algebra augmented with a special operator, called *best*, that can be used to specify queries based on qualitative preferences (Chomicki, 2003; Torlone and Ciaccia, 2002);

- h is an abstract hypertext definition expressed in WebML (Ceri et al, 2003), a conceptual model for Web application which which allows us to describe the structure of Web pages in a tight and concise way, by abstracting their logical features;

- p is presentation specification, expressed in terms of an original notion of *logical style sheet* that will be described in more detail below.

A Web page is composed by a set of *Web objects*, each of which is classified according to a predefined set of *Web Object Types (WOTs)*. Possible WOTs are *text*, *image*, *video*, *form* and so on. Each WOT τ is associated with a set of *presentation attributes*: they identify possible styles (e.g. font, color, spacing, position) that can be specified for τ. Given a set $W = \{w_1, \ldots, w_n\}$ of WOTs, a *logical style sheet* (or simply a *lls*) for W is a set of possible values for $w_1, \ldots, w_n$, respectively. Clearly, we can have different lss's for the same set of WOT's.

As an example, a graphical representation of a configuration $C = (q, h, p)$ is reported in Figure 2. In this figure, q has been represented by a tree, h by a site view (Ceri et al, 2003), and p by a table.

It is important to note that a configuration is indeed a logical notion that can be represented and implemented in several ways and with different syntaxes. This property guarantees the generality of the approach with respect to actual languages and tools used to implement the adaptive application. For instance, we can implement a configuration using SQL at the content level, XHTML at the navigation level and a set of CSS files at the presentation level.

We now introduce a composition operation over configurations, denoted by $\oplus$, that will be used in the adaptation process to merge different configurations. Given a pair configurations $C_1(q_1, h_1, p_1)$ and $C_2(q_2, h_2, p_2)$, $C_1 \oplus C_2$ is a configuration $C(q, h, p)$ defined as follows:

- $q = q_1 \circ q_2$, that is, q is obtained as the composition of q_1 followed by q_2;
- h is obtained by merging h_1 and h_2: if some conflict arises, the choices of h_1 are preferred to those of h_2; and
- $p(w_i) = p_1(w_i)$ if w_i is a WOT occurring in p_1 and $p(w_i) = p_2(w_i)$ otherwise (that is, if w_i is a WOT occurring only in p_2).

Note that the $\oplus$ operation is indeed a *prioritized* composition since if two configurations present conflicts, then the choices done in the configuration on the left hand side of $\oplus$ are preferred to the choices done in the other configuration.

Finally, a configuration can be associated with a profile to specify, in an abstract way, an adaptation suitable for the profile. In this case we say that the configuration *matches* the profiles. We do not give a precise definition of matching since, in general, this relationship is very difficult to establish in an automatic way. Therefore, we assume, at this stage, that the matching between profiles and configurations is predefined (e.g., specified by the designer of the application). We are currently investigating under which circumstances a match between a profile and a configuration can derived automatically.

3. A GENERAL METHODOLOGY OF ADAPTATION

3.1 The Adaptation Process

Our process of adaptation is based on the notions of profiles and configurations presented above. It relies on an initial set of configurations

C that capture the criteria of adaptations for a basic set of profiles. The only requirement is that **C** contains at least one configuration that match the generic profiles of a context. Clearly, this set needs to be refined and enriched during the life cycle of the adaptive system.

The methodology of adaptation can be summarized as follows:

1 First, the context of the client is captured and represented in terms of a set of profiles, one for each coordinate of adaptation. Each profile is expressed in the model presented in the previous section. As usual, some aspects of the context can be provided explicitly by the client (e.g., device capabilities), others can be specified implicitly (e.g., user preferences can be derived by the analysis of his/her navigation).

2 For each profile P of the context:

   (a) we select in **C** a configuration that matches with P;

   (b) if there is no configuration in **C** that matches P, we select in **C** a configuration that matches a profile P' that subsumes P.

3 The set of configurations generated in the previous step are merged into a unique configuration C making use of the $\oplus$ operator, on the basis of a predefined order of precedence on the various coordinates (e.g., the device capabilities take precedence over the preferences of the users).

4 The configuration C can be further refined to meet the requirements of a specific request done by the user (e.g., an additional data selection);

5 The final configuration C is translated into a corresponding set of *adaptation statements* that implement the configuration in the actual languages for the systems adopted at the various levels. Adaptation statements may correspond to: SQL statements at the content layer, XHTML or JSP statements at the navigation layer, and CCS style sheets at the presentation layer.

6 The adaptation statements are executed by the underlying systems and the final response is generated.

3.2 A Practical Example

Let us consider a data intensive Web site that publishes news items taken from different newspapers and assume that we intend to make it adaptive to different device characteristics and user preferences.

In Figure 3 is reported the whole process of adaptation when the site is accessed by a user having a preference of news and sport information with his/her cellular phone having a black and white display of limited size and without graphical capabilities.

First of all, the context of the client is captured and expressed in terms of a pair profiles P_d and P_u describing, respectively, the device and the user preferences. According to the technique presented in the previous section, two configurations are selected, one for each profile. The pair of configurations are then merged and we obtain in this way a final configuration representing, in an abstract way, the following adaptations.

- At the content level, the adaptation consists in a query that combines a projection introduced by P_d, which eliminates attributes that the device cannot display (e.g., pictures), and a selection introduced by P_u, which retrieves only news items preferred by the user, as indicated in the top of Figure 3.
- At the navigation level, the adaptation consists in the definition of an hypertext scheme that combines a distribution of contents into linked pages of limited size, to meet the requirements of P_d, and the separation of summaries and actual news, to meet the requirements of P_u; the scheme and the corresponding pages are reported in the middle of Figure 3.
- At the presentation level, the adaptation consists in the specification of black and white colors for P_d and of a font resize, as specified by the user in P_u; the logical style sheet we obtain is reported in the bottom of Figure 3 together with the corresponding output.

Note that the "logical" adaptations specified in the final configuration needs to be translated into actual specifications that can be then executed over the various levels to generate the desired interface.

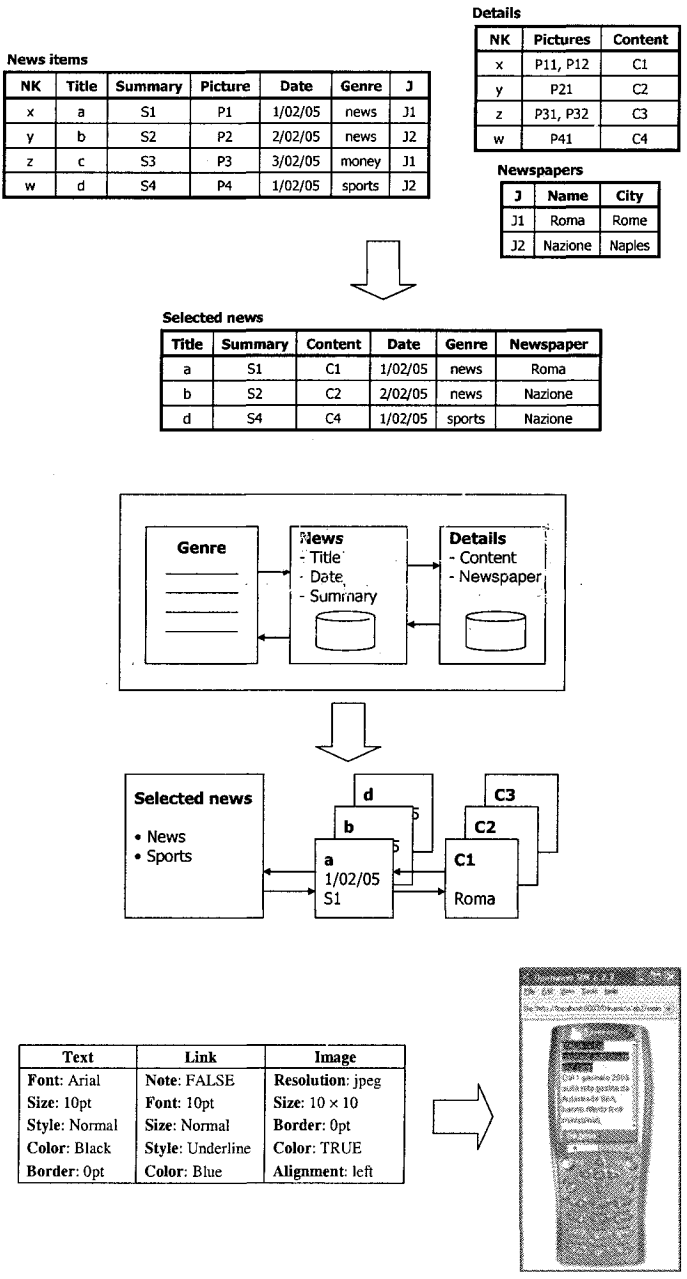
4. IMPLEMENTATION

4.1 An Architecture of Reference

A general architecture of a system able to meet the requirements of adaptation described in the Introduction and to implement the proposed methodology of adaptation is reported in Figure 4.

This includes:

- a Request Interpreter (RI), able to translate a specific user request (a page or a specific object) into a query over the underlying data,



News items

NK	Title	Summary	Picture	Date	Genre	J
x	a	S1	P1	1/02/05	news	J1
y	b	S2	P2	2/02/05	news	J2
z	c	S3	P3	3/02/05	money	J1
w	d	S4	P4	1/02/05	sports	J2

Details

NK	Pictures	Content
x	P11, P12	C1
y	P21	C2
z	P31, P32	C3
w	P41	C4

Newspapers

J	Name	City
J1	Roma	Rome
J2	Nazione	Naples

Selected news

Title	Summary	Content	Date	Genre	Newspaper
a	S1	C1	1/02/05	news	Roma
b	S2	C2	2/02/05	news	Nazione
d	S4	C4	1/02/05	sports	Nazione

Text	Link	Image
Font: Arial	**Note**: FALSE	**Resolution**: jpeg
Size: 10pt	**Font**: 10pt	**Size**: 10 × 10
Style: Normal	**Size**: Normal	**Border**: 0pt
Color: Black	**Style**: Underline	**Color**: TRUE
Border: 0pt	**Color**: Blue	**Alignment**: left

Figure 3. Adaptation at the various levels of a data intensive Web site

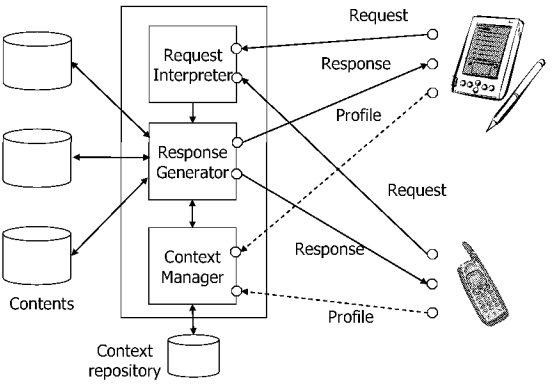


Figure 4. A general architecture of reference

- an Response Generator (RM), able to generate all the components of a response to deliver over the Web (content, structure and layout) that satisfies the given request and is appropriate for the client profile.
- a Context Manager (CM), able to get and manage a description of the client characteristics (the profiles) and support the Response Generator in the execution of its task.

Clearly, the fundamental component of this architecture is the Context Manager that should be able to:

1 (dynamically) capture and classify (possibly heterogeneous) incoming profiles of clients making use of a local repository of profiles,

2 coordinate the various (and possibly conflicting) requirements of adaptation for a given profile,

3 send to the Response Generator some *adaptation specifications* at all the levels (content, navigation and presentation) of the response.

To guarantee the flexibility of the overall system, this component should be extensible, in the sense that the various activities should be carried out for different types of profiles and according to orthogonal dimensions of adaptation, possibly not fixed in advance.

In Figure 5 it is reported a possible architecture for the Context Manager that can meet these requirements. The basic component of this

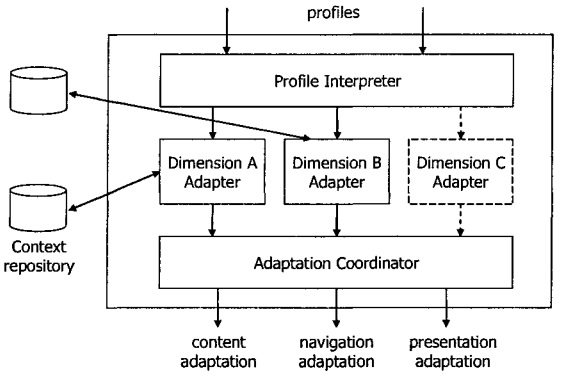


Figure 5. An Extensible Context Manager

module is the Profile Interpreter, which should be able to get and identify possibly heterogeneous profiles (e.g., CC/PP, XML, HTTP headers) and translate them into a uniform representation expressed in the model proposed in Section 1. Such profile representations are taken as input by a series of modules, one for each dimension of adaptation (e.g., the device characteristics, the user preferences, the location, etc.). The main task of these modules is to generate a uniform set of adaptation specifications, expressed in terms of a configurations, that satisfy the specific requirements of one dimension. This work can be supported by a special data repository in which predefined or previously generated profiles are stored together with their corresponding configurations. Since each module can generate different and possibly conflicting configurations, a coordination based on the $\oplus$ operator is performed to provide an integrated configuration that take into account the various adaptation requirements and can be effectively sent to the RG module. The Adaptation Coordinator is devoted to the execution of this task.

It is important to note that, due to the uniformity of representations and techniques used by the various adaptation modules, this scheme can be extended in a natural way: a new adaptation module can be easily added to satisfy the requirements of adaptation of a previously unpredicted coordinate.

A Response Generator that can match the other components of our architecture is composed by three modules (figure 6), one for each levels of the response to deliver over the Web. The first module combines the first component of the configuration provided by the Context Manager and generates from it a query to be executed by a Query Processor (possibly external to the system). The second module operates over the

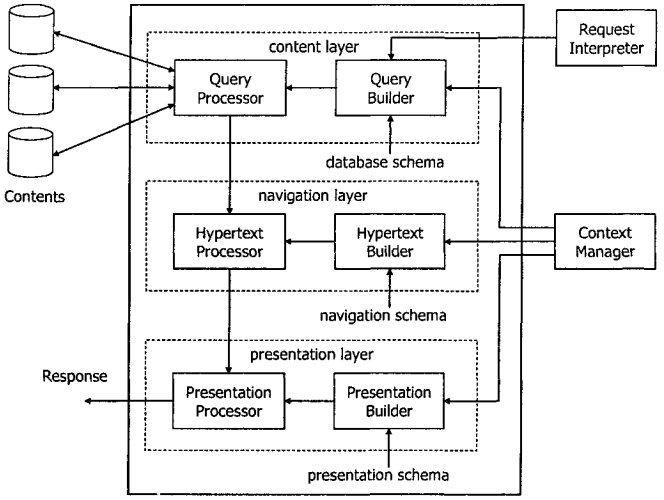


Figure 6. A Response Generator

navigation scheme of the Web site (e.g., by splitting pages or adding links) to satisfy the requirements of adaptation specified by the second component of the selected configuration. Finally, the last module is in charge of implementing the last component of the configuration at hand with an appropriate style sheet.

4.2 FAWIS

We have designed and developed the first release of a system called FAWIS (Flexible Adaptation of Web-based Information Systems) to test the our approach. The architecture of FAWIS is based on the general architecture reported in Figure 4 and also includes a Session Manager that is able to infer the interests of the user on a set of Web pages according the chronology of his/her navigation. The system relies on an XML database of contents and the selection of data is done by performing XML queries expressed in XQuery.

Currently, two modules of adaptation have been implemented, devoted to the management of the device capabilities and the user preferences, respectively. The repository for the former coordinate has been built by using the WURFL database (http://wurfl.sourceforge.net/), which includes information about capabilities and features of a large set of wireless devices currently available on the market. *Qualitative* preferences (Kießling, 2002; Torlone and Ciaccia, 2002) are used to represent the user profiles. The selection of contents based on qualitative

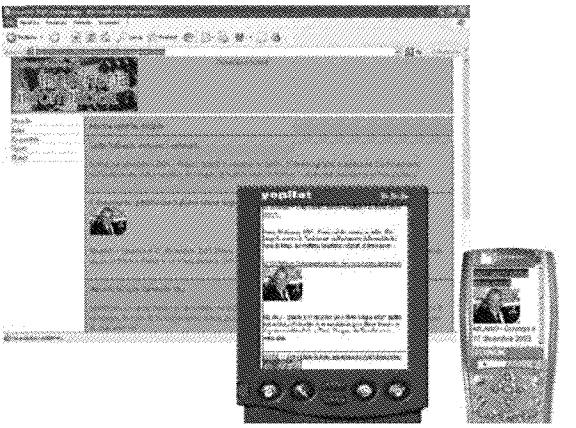


Figure 7. Three different results generated by FAWIS

preferences makes use of a special optimization technique that we have studied elsewhere (Torlone and Ciaccia, 2002).

In Figure 7 is reported the final results of the adaptation process performed by FAWIS, for the same information content, according to three different devices and user preferences.

In order to improve the efficiency of the system, we have implemented an optimization strategy based on the reuse of adaptation statements (see Section 3) and a cache of results. Statements defined for implementing a certain configuration are stored by the Response Generator in a special repository. When a new configuration C is generated by the Context Manager, the system verifies whether a suitable set of statements implementing C is already present in the repository. In this case, the system avoids the need for generating them again and efficiently executes the statements by taking advantage of a cache of already computed results.

An involved component of FAWIS is the profile interpreter that is currently able to interpret incoming profiles expressed in several languages including HTTP headers, UAProf (UAprof, 2001), CC/PP, and plain XML. These are translated into an homogeneous internal representation that refers to our profile model and is given as input to the various adaptation modules. We have equipped the Profile Interpreter with a graphical interface that allows the user to define new formats of profiles and specify mappings between the various formats and the general representation. These mappings are used to perform the translation of the incoming profiles.

We have tested our system to experiment the effectiveness and the efficiency of the approach illustrated in this paper in several practical cases.

On the server side, we used an IBM computer xSeries 225, equipped with a Xeon2 2.8Ghz processor, a 4 GB RAM, and a 120 GB HDD SCSI. The Web site have been accessed by three different types of devices: a mid-range desktop, several PDAs with different capabilities, and some cellular phones.

In general, the results obtained so far are promising. The effectiveness of the approach is supported by the capability demonstrated by the system to generate satisfactory results for different contexts, even not fixed in advance. This is due to the ability of the approach to select configurations for "compatible" context, as described in Section 3. Also, it turned out that the adaptation process can be executed always in a reasonable amount of time and that the response time rapidly improves when the caching of adaptation statements becomes effective.

5. CONCLUSION AND FURTHER WORK

In this paper, we have presented a general approach to the problem of content delivery adaptation of Web information. The approach is based on a general notion of profile that can be used to represent a variety of contexts at different level of details. Each profile is associated with a configuration that specifies, in abstract terms, how the response for a given request has to be delivered over the Web, by taking into account the requirements of adaptation for the profile. We have presented a general methodology for content adaptation based on the generation and management of configurations. We have also described and tested a prototype implementing the proposed approach.

From a conceptual point of view, we are currently investigating in more depth the notions of profile and configuration, in order to improve their generality and usability. From a practical point of view we are extending the features of the tool, by enhancing the profile interpretation capabilities and by adding new adaptation modules.

References

T. Bickmore, A. Girgensohn, and J. Sullivan. Web page filtering and reauthoring for mobile users. *Computer Journal 42(6)* (1999), 534–546.

D. Brickley and R. Guha. *RDF vocabulary Description Language 1.0: RDF Schema.* W3C Working Draft, 10 October 2003.

S. Ceri, P. Fraternali, and A. Bongio. Web modeling language (webml): a modeling language for designing web sites. In *9th International Conference on the World Wide Web (WWW9) Amsterdam*, 2000.

S. Ceri, P. Fraternali, A. Bongio , M. Brambilla, S. Comai, and M. Matera. *Designing Data-Intensive Web Applications.* Morgan Kaufmann, 2003.

J. Chomicki. Preference formulas in relational queries. *ACM Trans. Database Syst. 28(4)*: 427–466, 2003.

Z. Fiala, M. Hinz, K. Meissner, and F. Wehner. A component-based approach for adaptive dynamic web documents. *Journal of Web Engineering, Rinton Press 2*:058–073, 2003.

Z. Fiala , F. Frasincar , M. Hinz, G.J. Houben, P. Barna, K. Meißner. Engineering the presentation layer of adaptable web information systems. In *International Conference on Web Engineering (ICWE 2004) Glasgow, Scotland*, (July 28-30 2004).

F. Frasincar, P. Barna, G.J. Houben, and Z. Fiala. Adaptation and reuse in designing web information systems. In *International Conference on Information Technology, Track on Modern Web and Grid Systems, (IEEE Compute Society 2004) Glasgow, Scotland* (387–391).

W. Kießling. Foundations of preferences in database systems. In *(VLDB 2002) pp. 311-322* (2002).

G. Klyne, F. Reynolds, C. Woodrow, H. Ohto, J. Hjelm, M. Butler, and L. Tran. *Composite Capability/Preference Profiles (CC/PP): Structure and Vocabularies.* W3C Working Draft, 2004.

W. Gu and A. S. Helal. An XML Based Solution to Delivering Adaptive Web Content for Mobile Clients. In *International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS'04), San Jose, California, July 25–29* (2004).

S. Gupta, G. Kaiser,, D. Neistadt, and P. Grimm. Dom-based content extraction of html documents. In *Twelfth International Conference on the World Wide Web (WWW2003), Budapest, Hungary* (2003).

O. Lassila and R. Swick. *Resource description framework (rdf) model and syntax specification.* W3C Working Draft, 22 February 1999.

A. Leubner and W. Kießling. Personalized keyword search with partial-order preferences. In (SBBD 2002: 181-193), 2002.

Open Mobile Alliance WAP Forum 2001. *Wireless Application Group: User Agent Profile Specification.*

O. Pastor, J. Fons, and V. Pelechano. A method to develop web applications from web-oriented conceptual models. In *International Workshop on Web Oriented Software Technology (IWWOST)* ((2003) 144–173).

D. Schwabe, G. Rossi, and S.D.J. Barbarosa. Systematic hypermedia application design with oohdm. In *Hypertext '96, The Seventh ACM Conference on Hypertext ACM Washington DC* (1996), 116–128.

R. Torlone and P. Ciaccia. Which are my Preferred Items? In *2nd Int. Workshop on Recommendation and Personalization in e-Commerce, Malaga, Spain*, pag. 1–9, 2002.

R. Vdovjak, F. Fransincar, G.J. Houben, and P. Barna. Engineering semantic web information systems in hera. *Journal of Web Engineering, Rinton Press 2* (2003), 003–026.

M. Wagner, W. Balke, W. Kießling. An xml-based multimedia middleware for mobile online auctions. In *(ICEIS(2)2001: 934-944)* (2001).

W3C Working Group on Device Independence. Device Independence Principles. *Internet document: http://www.w3.org/TR/di-princ/*, 2003.

DISTRIBUTED CONTEXT MONITORING FOR CONTINUOUS MOBILE SERVICES

Claudio Bettini, Dario Maggiorini, and Daniele Riboni
DICo, University of Milan, via Comelico 39, I-20135, Milan, Italy

Abstract: Context-awareness has been recognized as a very desirable feature for mobile internet services. This paper considers the acquisition of context information for continuous services, i.e., services that persist in time, like streaming services. Supporting context-awareness for these services requires the continuous monitoring of context information. The paper presents the extension of a middleware architecture for the reconciliation of distributed context information to support context-aware continuous services. The paper also addresses optimization issues and illustrates an adaptive video streaming prototype used to test the middleware.

Keywords: Context-awareness, adaptation, continuous mobile services

1. INTRODUCTION

Internet services provided to mobile users can be divided into two broad categories: instantaneous (or one-shot) services and continuous services. Examples of services in the first category are web browsing, search for the closest pharmacy, and delivery of a message with the current balance of a certain bank account. Services in the second category persist much longer in time and are typically characterized by multiple transmissions of data by the service provider. Examples are multimedia streaming, navigation services, location-based recommendation services, and publish/subscribe services.

Services in the first category can be implemented with context-aware features if some context information can be obtained at the time of service request and if the service application logic can take this information into account when answering to a specific request. Context-awareness is much

more challenging for continuous services, since changes in context should be taken into account during service provisioning. As an example, consider an adaptive streaming service. Typically, parameters used to determine the most appropriate media quality include a number of context parameters, as, for example, an estimate of the available bandwidth and the battery level on the user's device. Note that this information may be owned by different entities, e.g., the network operator and the user's device, respectively. With a naïve approach, the application logic should constantly monitor these parameters, possibly by polling servers in the network operator's infrastructure as well as the user's device for parameter value updates. Moreover, the application logic should internally re-evaluate the rules that determine the streaming bit rate (e.g., "if the user's device is low on memory, decrease the bitrate"). This approach has a number of shortcomings, including: (*i*) client-side resource consumption; (*ii*) high response times due to the polling strategy; (*iii*) complexity of the application logic; (*iv*) poor scalability, since for every user the service provider must continuously request context information and re-evaluate its rules. An alternative approach is to provide the application logic with asynchronous notifications of relevant context changes, on the basis of its specific requirements. However, when context information must be aggregated from distributed sources which may possibly deliver conflicting values, as well as provide different dependency rules among context parameters, the management of asynchronous notifications is far from trivial. The **CARE** middleware (Agostini et al., 2004; Bettini and Riboni, 2004) was originally designed to support instantaneous context-aware mobile services in an environment characterized by distributed context sources. The main contribution of this paper is indeed the extension of the **CARE** middleware to include a mechanism of asynchronous notifications, enabling context-awareness also for continuous mobile services. Technically, the extension involves algorithms to identify context sources and specific context parameter thresholds for these sources, with the goal of minimizing the exchange of data through the network and the time to re-evaluate the rules that lead to the aggregated context description.

While several frameworks have been proposed to support context awareness (see e.g., Bellavista et al., 2003; Butler et al., 2002; Chen et al., 2004; Hull et al., 2004), we have extensively described elsewhere (Agostini et al., 2004) that **CARE** has some unique features in dealing with distributed and possibly conflicting context information. The extension to the support of asynchronous context change notifications still preserves these unique features. The work on stream data management has probably the closer connection with the specific problem we are tackling. Indeed, each source of context data can be seen as providing a stream of data for each context parameter it handles. One of the research issues considered in that area is the

design of filter bound assignment protocols with the objective of reducing communication cost (see, e.g., Chengy et al., 2005). Since *filters* are the analogous of *triggers* used in our approach to issue asynchronous notifications, we are investigating the applicability of some of the ideas in that field to our problem.

The rest of the paper is organized as follows: Section 2 describes the **CARE** middleware architecture and features; Section 3 contains the principles and technical details for managing asynchronous notifications; Section 4 shows how the extension has been implemented; Section 5 briefly describes a streaming service used to test the middleware; Section 6 concludes the paper.

2. THE MIDDLEWARE ARCHITECTURE

The **CARE** (Context Aggregation and REasoning) middleware has been presented in detail elsewhere (Agostini et al., 2004; Bettini and Riboni, 2004). Here we only describe what is needed to understand the extension to support continuous services.

2.1 Overview

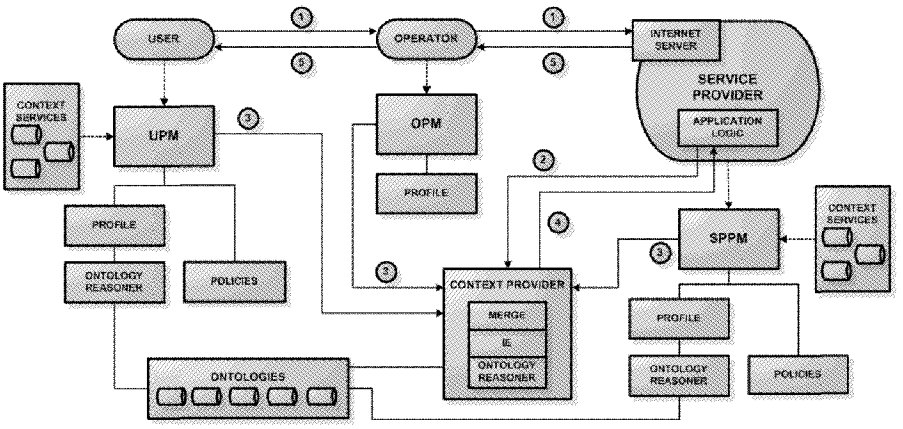


Figure 1. Architecture overview and data flow upon a user request

In our middleware, three main entities are involved in the task of building an aggregated view of context information, namely: the *user* with his/her devices, the *network operator* with its infrastructure, and the *service*

provider with its own infrastructure. Clearly, the architecture has been designed to handle an arbitrary number of entities. In **CARE** we use the term *profile* to indicate a set of context parameters, and a profile manager is associated with each entity; profile managers are named *user profile manager* (UPM), *operator profile manager* (OPM), and *service provider profile manager* (SPPM), for the user, the network operator and the service provider, respectively. Adaptation and personalization parameters are determined by policy rules defined by both the user and the service provider, and managed by their corresponding profile managers. In Figure 1 we illustrate the system behavior by describing the main steps involved in a service request. At first (step 1) a user issues a request to a service provider through his device and the connectivity offered by a network operator. The HTTP header of the request includes the URIs of UPM and the OPM. Then (step 2), the service provider forwards this information to the CONTEXT PROVIDER asking for the profile information needed to perform adaptation. In step 3, the same module queries the profile managers to retrieve distributed profile data and user's policies. Profile data are aggregated by the MERGE module in a single profile which is given, together with policies, to the Inference Engine (IE) for policy evaluation. In step 4, the aggregated profile is returned to the service provider. Finally, profile data are used by the application logic to properly adapt the service before its provision (step 5). Our architecture can also interact with ontological reasoners, but this aspect will not be addressed in this paper.

2.2 Profile Aggregation

In the following we show how possibly conflicting data can be aggregated into a single profile.

Profile and policy representation

Essentially, profiles are represented adopting the CC/PP (Klyne et al., 2004) specification, and can possibly contain references to ontological classes and relations. However, for the sake of this paper, we can consider profiles as sets of *attribute/value* pairs. Each attribute semantics is defined in a proper vocabulary, and its value can be either a *single value*, or a *set/sequence* of single values.

Policies are logical rules that determine the value of profile attributes on the basis of the values of other profile attributes. Hence, each policy rule can be interpreted as a set of conditions on profile data that determine a new value for a profile attribute when satisfied.

Example 1 *Consider the case of a streaming service, which determines the most suitable media quality on the basis of network conditions and available memory on the user's device. The* MediaQuality *is determined by the evaluation of the following policy rules:*

R1: *"**If*** AvBandwidth ≥ *128kbps **And*** Bearer = *'UMTS' **Then Set*** NetSpeed= *'high'"*
R2: *"**If*** NetSpeed= *'high' **And*** AvMem ≥ *4MB **Then Set*** MediaQuality= *'high'"*
R3: *"**If*** NetSpeed= *'high' **And*** AvMem < *4MB **Then Set*** MediaQuality= *'medium'"*
R4: *"**If*** NetSpeed! = *'high' **Then Set*** MediaQuality= *'low'"*

*Rules R2, R3 and R4 determine the most suitable media quality considering network conditions (*NetSpeed*) and available memory on the device (*AvMem*). In turn, the value of the* NetSpeed *attribute is determined by rule R1 on the basis of the current available bandwidth (*AvBandwidth*) and* Bearer.

Conflict resolution

We recall that, once the CONTEXT PROVIDER has obtained profile data from the other profile managers, at first this information is passed to the MERGE module which is in charge of merging profiles. Conflicts can arise when different values are provided by different profile managers for the same attribute. For example, suppose that OPM provides for the *AvBandwidth* attribute a certain value *x*, while the SPPM provides for the same attribute a different value *y*, obtained through some probing technique. In order to resolve this type of conflict, the CONTEXT PROVIDER has to apply a resolution rule at the attribute level. These rules (called *profile resolution directives*) are expressed in the form of priorities among entities, which associate to every attribute an ordered list of profile managers.

Example 2 *Consider the following profile resolution directives, set by the provider of the streaming service cited in Example 1:*

PRD1: setPriority AvBandwidth = (OPM, SPPM, UPM)
PRD2: setPriority MediaQuality = (SPPM, UPM)

In PRD1, the service provider gives highest priority to the network operator for the AvBandwidth *attribute, followed by the service provider and by the user. The absence of a profile manager in a directive (e.g., the absence of the OPM in PRD2) states that values for that attribute provided by that profile manager should never be used. The conflict described above is resolved by applying PRD1. In this case, the value x is chosen for the*

available bandwidth. The value y would be chosen in case the OPM does not provide a value for that attribute.

Once conflicts between attribute values provided by different profile managers are solved, the resulting merged profile is used for evaluating policy rules. Since policies can dynamically change the value of an attribute that may have an explicit value in a profile, or that may be changed by some other policies, they introduce nontrivial conflicts. The intuitive strategy is to assign priorities to rules having the same head predicate on the basis of its *profile resolution directive*. Hence, rules declared by the first entity in the *profile resolution directive* have higher priority with respect to rules declared by the second entity, and so on. When an entity declares more than one rule with the same head predicate, priorities are applied considering the explicit priorities given by that entity.

Example 3 *Consider the set of rules shown in Example 1 and* profile resolution directives *shown in Example 2. Suppose that R2 and R3 are declared by the service provider, and R4 is declared by the user. Since the service provider declared two rules with the same attribute in the head, it has to declare an explicit priority between R2 and R3. Suppose the service provider gives higher priority to R2 with respect to R3. Since the SPPM has higher priority with respect to the UPM, according to the* profile resolution directive *regarding* MediaQuality *(i.e., PRD2), if p(R) is the priority of rule R, we have that:*

$$p(R2)>p(R3)>p(R4)$$

The intuitive evaluation strategy is to proceed, for each attribute A, starting from the rule having $A()$ in its head with the highest priority, and continuing considering rules on $A()$ with decreasing priorities till one of them fires. If none of them fires, the value of A is the one obtained by the MERGE module on A, or *null* if such a value does not exist. A more in-depth discussion of conflict resolution can be found in (Bettini and Riboni, 2004).

3. SUPPORTING CONTINUOUS MOBILE SERVICES

In this section we describe a trigger mechanism for supporting continuous mobile services (Maggiorini and Riboni, 2005). This mechanism allows profile managers to asynchronously notifying the service provider upon relevant changes in profile data on the basis of triggers. Triggers in this

case are essentially conditions over changes in profile data (e.g., available bandwidth dropping below a certain threshold, or a change of the user's activity) which determine the delivery of a notification when met. In particular, when a trigger fires, the corresponding profile manager sends the new values of the modified attributes to the CONTEXT PROVIDER module, which should then re-evaluate policies.

3.1 Trigger-based Mechanism

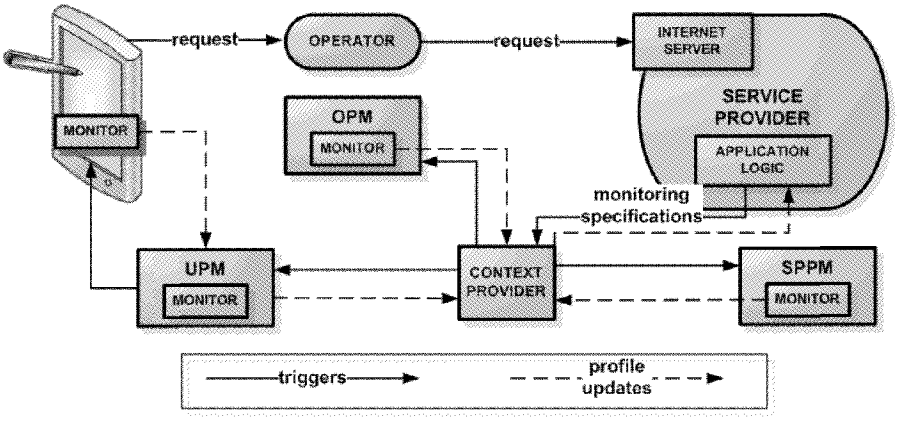


Figure 2. Trigger mechanism

Figure 2 shows an overview of the mechanism. To ensure that only useful update information is sent to the service provider, a deep knowledge of the service characteristics and requirements is needed. Hence, the context parameters and associated threshold values that are relevant for the adaptation (named *monitoring specifications*) are set by the service provider application logic, and communicated to the CONTEXT PROVIDER. Actual triggers are generated by the CONTEXT PROVIDER –according to the algorithms presented in the following of this section– and communicated to the proper profiles managers. Since most of the events monitored by triggers sent to the UPM are generated by the user device, the UPM communicates triggers to a light server module resident on the user's device. Note that, in order to keep up-to-date the information owned by the UPM, each user device must be equipped with an application monitoring the state of the device against the received triggers (named MONITOR in Figure 2), and with an application that updates the UPM when a trigger fires. Each time a profile manager receives an update for a profile attribute value that makes a trigger

fire, it forwards the update to the CONTEXT PROVIDER. Then, the CONTEXT PROVIDER re-computes the aggregated profile, and any change satisfying a monitoring specification is communicated to the application logic. In order to show the system behavior, consider the following example.

Example 4 *Consider the case of the streaming video service introduced in Example 1. Suppose that a user connects to this service via a UMTS connection, and that at first the available bandwidth is higher than 128kbps, and the user device has more than 4MB available memory. Thus, the CONTEXT PROVIDER, evaluating the service provider policies, determines a high* MediaQuality *(since rules R1 and R2 fire). Consequently, the service provider starts the video provision with a high bitrate. At the same time, the application logic sets a* monitoring specification *regarding* MediaQuality. *Analyzing policies, profile resolution directives, and context data, the CONTEXT PROVIDER sets triggers to the OPM and to the UPM/device, asking a notification in case the available bandwidth and the available memory, respectively, drop below certain thresholds. Suppose that, during the video provision, the user device runs out of memory. Then, the UPM/device sends a notification (together with the new value for the available memory) to the CONTEXT PROVIDER, which merges profiles and re-evaluates policies. Since this time policy evaluation determines a lower* MediaQuality *(since rule R3 fires), the video bitrate is immediately lowered by the application logic.*

3.2 Monitoring Specifications

In order to keep the re-evaluation of rules to a minimum, it is important to let the application logic to precisely specify the changes in context data it needs to be aware of in order to adapt the service. These adaptation needs, called *monitoring specifications*, are expressed as conditions over changes in profile attributes. As an example, consider the provider of the continuous streaming service shown in Example 1. The application logic only needs to be aware of changes to be applied to the quality of media. Hence, its only *monitoring specification* will be:

$$MediaQuality(X),\ X \neq \$old\_value\_MediaQuality,$$

where *$old_value_MediaQuality* is a variable to be replaced with the value for the *MediaQuality* attribute, as retrieved from the aggregated profile. *Monitoring specifications* are expressed through an extension of the language used to define rule preconditions in our logic programming language (Bettini and Riboni, 2004). This extension involves the introduction of the additional special predicate *difference*, which has the

obvious semantics with respect to various domains, including spatial, temporal, and arithmetic domains. For instance, the *monitoring specification*:

Coordinates(*X*), *difference*(*X*, *$old_value_Coordinates*) > 200 *meters*

will instruct the CONTEXT PROVIDER to notify changes of the user position greater than 200 meters.

3.3 Minimizing Unnecessary Updates

In general, allowing the application logic to specify the changes in context data it is interested to does not prevent that unnecessary updates are sent to the CONTEXT PROVIDER. We define an update to the value of a profile attribute as *unnecessary* if it does not affect the aggregated profile. In the context of mobile service provisioning, the cost of unnecessary updates is high, in terms of client-side bandwidth consumption (since updates can be sent by the user's device), and server-side computation, and can compromise the scalability of the architecture. In order to avoid inefficiencies, the application logic does not directly communicate *monitoring specifications* to the profile managers. Instead, *monitoring specifications* are communicated to the CONTEXT PROVIDER, which is in charge of deriving the actual triggers and performing the optimizations that will be described in the following of this section.

Baseline algorithm

The baseline algorithm for trigger derivation consists of the following steps: *a)* set a trigger regarding the attribute A_i for each *monitoring specification* c_{A_i} regarding A_i, *b)* communicate the trigger to every profile manager, and *c)* repeat this procedure considering each precondition of the rules having A_i in their head as a monitoring specification. As a matter of fact, if A_i is an attribute whose value can be possibly modified by policies (i.e., an attribute that appears in the head of some policy rule), it is not sufficient to monitor the single A_i attribute. For instance, consider rule *R*2 in Example 1. The value of the *MediaQuality* attribute depends on the values of other attributes, namely *NetSpeed* and *AvMem*. Hence, those attributes must also be kept up-to-date in order to satisfy a *monitoring specification* regarding *MediaQuality*. For this reason, the CONTEXT PROVIDER sets new triggers regarding those attributes. Note that this mechanism must be recursively repeated accordingly to step *c)*. For example, since *NetSpeed*

depends on *AvBandwidth* and *Bearer*, the CONTEXT PROVIDER would set two triggers regarding those attributes. Generally speaking, for each *monitoring specification* c_{A_i} regarding an attribute A_i whose value was set by rule r'_{A_i}, triggers must be set for checking that the preconditions of rule r'_{A_i} are still valid, and for monitoring the preconditions of the other rules that can set a value for A_i.

The use of the baseline algorithm would lead to a number of unnecessary updates, as will be shortly explained in Example 5. We devised two optimizations, presented in the following of this section, which avoid a large number of unnecessary updates while preserving useful ones.

Optimization based on profile resolution directives

Any update that does not affect the profile obtained after the MERGE operation is unnecessary. Indeed, since we assume that neither policies, nor *profile resolution directives* can change during service provision, the aggregated profile does not change as long as the profile obtained after the *merge* operation remains the same. Hence, the first optimization considers the profile resolution directives used by the *merge* operation. The semantics of *merge* ensures that the value provided by an entity e_i for the attribute a_j can be overwritten only by values provided by e_i or provided by entities which have higher priority for the a_j attribute.

Example 5 *Consider the* profile resolution directive *on the attribute* AvBandwidth *given in Example 2 (PRD1). Suppose that the OPM (the entity with the highest priority) does not provide a value for* AvBandwidth, *but the SPPM and the UPM do. The value provided by the SPPM is the one that will be chosen by the MERGE module, since the SPPM has higher priority for that attribute. In this case, possible updates sent by entities with lower priority than the SPPM (namely, the UPM) would not modify the profile obtained after the MERGE operation, since they would be discarded by the merge algorithm. As a consequence, the CONTEXT PROVIDER does not communicate a trigger regarding* AvBandwidth *to the UPM.*

Note that, if the application logic defines a *monitoring specification* regarding an attribute whose value is *null* (i.e., an attribute for which no profile manager provided a value), the corresponding trigger is communicated to every entity that appears in the *profile resolution directive.*

Optimization based on rule priority

The second optimization exploits the fact that an attribute value set by the rule r'_{A_i} can be overwritten only by r'_{A_i} or by a rule having higher priority than r'_{A_i}. As a consequence, values set by rules having lower priority than r'_{A_i} are discarded, and do not modify the aggregated profile. For this reason, the preconditions of rules r_{A_i} having lower priority with respect to r'_{A_i} should not be monitored.

Generally speaking, for each *monitoring specification* c_{A_i}, an implicit *monitoring specification* is created for each precondition of the rule r'_{A_i} that determined the last value for A_i, and for the preconditions of the other rules having A_i in their head, and having higher priority than r'_{A_i}. Rules with lower priority do not generate triggers. For each *monitoring specification*, the CONTEXT PROVIDER creates a trigger and communicates it to the proper profile managers, as explained in Section 3.3.2.

Example 6 *Consider rules R2, R3 and R4 in Example 1. We recall from Example 3 that p(R2)>p(R3)>p(R4), where p(R) is the priority of rule R. Rules are evaluated in decreasing order of priority. Suppose that R2 does not fire, while R3 fires. In this case, the preconditions of R2 (the only rule with higher priority in this example) must be monitored, since they can possibly determine the firing of this rule. Preconditions of R4 must not be monitored since, even if they are satisfied, R4 cannot fire as long as the preconditions of R3 are satisfied. The preconditions of R3 must be monitored in order to assure that the value derived by the rule is still valid. In case the preconditions of R3 do not hold anymore, rules with lower priority (R4 in this example) can fire, and their preconditions are added to the set of implicit* monitoring specifications.

4. SOFTWARE ARCHITECTURE

The software architecture used to implement our middleware is shown in Figure 3. We have chosen Java as the preferred programming language, switching to more efficient solutions only when imposed by efficiency requirements. With regard to the inter-modules communication we have preferred, where possible, the web service paradigm.

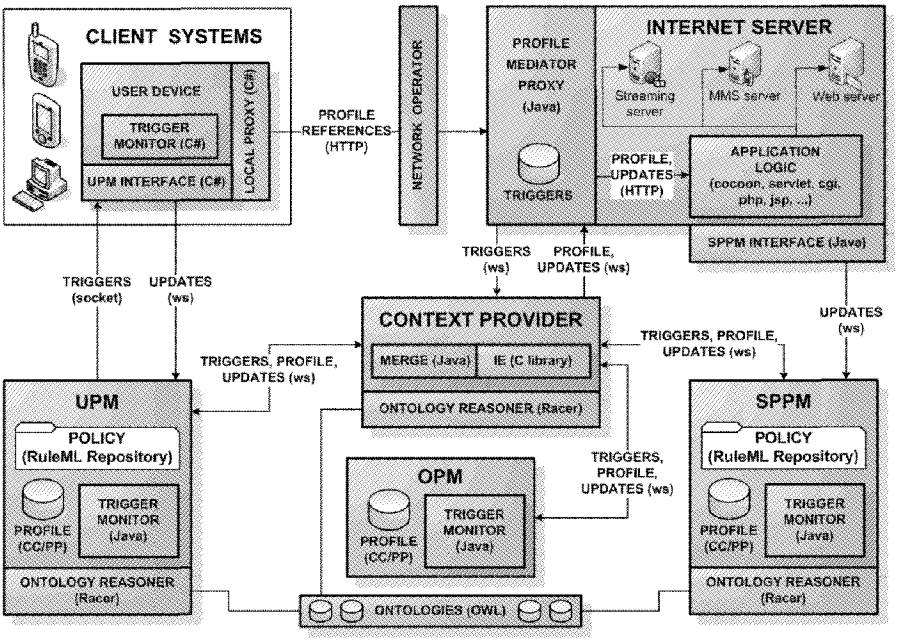


Figure 3. The software architecture

The PROFILE MEDIATOR PROXY (PMP) is a server-side Java proxy that is in charge of intercepting the HTTP requests from the user's device, and of communicating the user's profile (retrieved from the CONTEXT PROVIDER) to the application logic, by inserting profile data into the HTTP request headers. In this way, user profile data is immediately available to the application logic, which is relieved from the burden of asking the profile to the CONTEXT PROVIDER, and of parsing CC/PP data. The PMP is also in charge of storing the *monitoring specifications* of the application logic. When the PMP receives a notification of changes in profile data, it communicates them to the application logic by means of an HTTP message. Given the current implementation of the PMP, the application logic can be developed using any technology capable of parsing HTTP requests, including JSP, PHP, Cocoon, Java servlets, ASP .NET, and many others. The application logic can also interact with provisioning servers based on protocols other than HTTP. For instance, in the case of the adaptive streaming server presented in Section 5, profile data are communicated to the streamer by a PHP script through a socket-based protocol.

CC/PP parsing is performed using RDQL, a query language for RDF documents implemented by the Jena Toolkit[1] . User and service provider policies are represented in RuleML (Boley et al., 2001). The evaluation of the logic program is performed by an efficient, ad-hoc inference engine developed using C.

Profile data, policies and triggers are stored by the profile managers into ad-hoc repositories that make use of the MySQL DBMS. Each time a profile manager receives an update of profile data, the TRIGGER MONITOR evaluates the received triggers, possibly notifying changes to the CONTEXT PROVIDER. The UPM has some additional modules for communicating triggers to a server application executed by the user device. The communication of triggers is based on a socket protocol, since the execution of a SOAP server by some resource-constrained devices could be unfeasible.

The TRIGGER MONITOR module on the user's device is in charge of monitoring the status of the device (e.g., the battery level and available memory) against the received triggers. The LOCAL PROXY is the application that adds custom fields to the HTTP request headers, thus providing the CONTEXT PROVIDER with the user's identification, and with the URIs of his UPM and OPM. At the time of writing, modules executed on the user device are developed using C# for the .NET (Compact) Framework.

5. AN ADAPTIVE MULTIMEDIA STREAMING SERVICE

As already discussed in previous work (Maggiorini and Riboni, 2005), multimedia streaming adaptation can benefit from an asynchronous messaging middleware. In order to demonstrate the effectiveness of our solution, we implemented a streamer prototype based on the middleware described in this paper. We chose the VideoLan Client (*VLC*)[2] as a starting point to develop a customized client system, because it is an open platform and multiple operating systems are supported. The client is intended to run on windows workstations and windowsCE PDAs in order to achieve the largest possible population of users. VLC contacts the streaming service provider performing an HTTP request that has been modified by a local proxy adding in the HTTP headers the URIs of UPM and OPM. Then, the client waits for the video feed to come on a specific port. The HTTP request from VLC is received by the PMP module, which, as explained in Section 4, asks the CONTEXT PROVIDER for the aggregated profile information. The

[1] `http://jena.sourceforge.net/`
[2] `http://www.videolan.org/`

returned attribute/value pairs are included in the HTTP request header and the request is forwarded to the streamer application logic. Upon receiving the request, the streamer opens all the video files with the different encodings. Based on the context parameter values, the application logic selects an appropriate encoding, and the streamer starts sending over the network UDP packets containing frames belonging to the selected encoding. The streamer has been implemented on a Linux system. Network streaming is performed thanks to a specific file format, in which video data is already divided in packets, and a network timestamp is associated to each packet. Moreover, this streaming file format supports any kind of encoding, thus making the service independent by any specific format or codec.

We now illustrate how changes in context are detected and notified by the middleware to the streamer application logic. When the PMP module receives the user request, it recognizes that is directed to a continuous service, and retrieves from the media description all the monitoring specifications related to the requested feed, which in the case of our streamer prototype consist only of the *MediaQuality* parameter. Using the specification, the CONTEXT PROVIDER computes the set of required triggers, accordingly to the algorithms reported in Section 3, and illustrated by Example 6. Triggers are then set on the OPM, UPM/device to monitor available bandwidth and battery level, respectively. Upon firing of one of the triggers, the new value is forwarded to the CONTEXT PROVIDER, which recomputes the value for the *MediaQuality* parameter. If the new value differs from the previous one, it is forwarded to the PMP which issues a special HTTP request to the streamer application logic. The application logic selects a different encoding based on the new value; the feeder process is notified and forced to change the file from which the video frames are taken. The preliminary experiments performed with the current prototype are based on a full implementation and demonstrate the viability of our solution.

6. CONCLUSIONS

We presented the extension of the **CARE** middleware to support context-aware continuous services. In particular, we focused on optimizations to avoid unnecessary remote context updates which would strongly affect scalability. An adaptive video streamer has been used for a practical evaluation of the functionality of our middleware.

The natural extension of the **CARE** architecture will be related to session handoff. We experience a session handoff every time a user switches device and/or network connection. The main issue here is to efficiently describe all context data regarding the current session in order to guarantee a seamless

migration to the new device. In the case of our video streaming prototype, session handoff will be easy to accomplish, since session data can be easily described as the current frame number and media quality. In a more general scenario, involving a possibly huge number of context data, session handoff is much more challenging, and is the subject of future work.

References

Agostini, A., Bettini, C., Cesa-Bianchi, N., Maggiorini, D., Riboni, D., Ruberl, M., Sala, C., and Vitali, D., 2004, Towards highly adaptive services for mobile computing, *Proc. of IFIP TC8 Working Conference on Mobile Information Systems (MOBIS)*, Springer, pp. 121–134.

Bellavista, P., Corradi, A., Montanari, R., and Stefanelli, C, 2003, Context-aware middleware for resource management in the wireless Internet, *IEEE Trans. Soft. Eng., Special Issue on Wireless Internet*, **29**(12):1086–1099.

Bettini, C., and Riboni, D., 2004, Profile aggregation and policy evaluation for adaptive Internet services, *Proc. of the First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (Mobiquitous)*, IEEE, pp. 290–298.

Boley, H., Tabet, S., and Wagner, G., 2001, Design rationale of RuleML: a markup language for Semantic Web rules, *Proc. of the International Semantic Web Working Symposium (SWWS)*, pp. 381–401.

Butler, M., Giannetti, F., Gimson, R., and Wiley, T., 2002, Device independence and the Web, *IEEE Internet Comp.*, **6**(5):81–86.

Chen, H., Finin, T., and Joshi, A., 2004, Semantic Web in the context broker architecture, *Proc. of the Second IEEE International Conference on Pervasive Computing and Communications (PerCom 2004)*, IEEE, pp. 277–286.

Chengy, R., Kaox, B., Prabhakary, S., Kwanx, A., and Tu, Y., 2005, Adaptive stream filters for entity-based queries with non-value tolerance, *Proc. of VLDB 2005, International Conference on Very Large Databases*.

Hull, R., Kumar, B., Lieuwen, D., Patel-Schneider, P., Sahuguet, A., Varadarajan, S., and Vyas, A., 2004, Enabling context-aware and privacy-conscius user data sharing, *Proc. of the 2004 International Conference on Mobile Data Management*, IEEE, pp. 187–198.

Klyne, G., Reynolds, F., Woodrow, C., Ohto, H., Hjelm, J., Butler, M. H., and Tran, L., 2004, Composite capability/preference profiles (CC/PP): structure and vocabularies 1.0. W3C recommendation, http://www.w3.org/TR/2004/REC-CCPP-struct-vocab-20040115/.

Maggiorini, D., and Riboni, D., 2005, Continuous media adaptation for mobile computing using coarse-grained asynchronous notifications, *2005 International Symposium on Applications and the Internet (SAINT 2005), Proc. of the Workshops*, IEEE, pp. 162–165.

MOBILE-WEB SERVICES VIA PROGRAMMABLE PROXIES*

R. Grieco[1], D. Malandrino[1], F. Mazzoni[2], and V. Scarano[1]

[1] *ISISLab, Dipartimento di Informatica ed Applicazioni "R.M. Capocelli" Università di Salerno, 84081 Baronissi (Salerno), Italy*

[2] *Dipartimento di Ingegneria dell'Informazione Università di Modena e Reggio Emilia, 41100 Modena, Italy*

Abstract: Our goal, in this paper, is to present the effectiveness of an intermediary framework to provide mobile-oriented services via edge services. To this end we developed services for device independence in such a way that content is adapted according to the capabilities of the target devices.

Keywords: Programmable proxies, device independence, ubiquitous Web.

1. INTRODUCTION

The World Wide Web is, nowadays, a ubiquitous infrastructure for any information and communication technology. In fact, it is globally recognized as a universal framework to *anytime-anywhere* access *any data* or information from any networked computer or device and by using *any access network*.

The proliferation of the rich, entertaining and interactive applications available on the Web is mimicked by the comparable growth in wireless technologies that spawn a variety of terminal devices such as Desktop PC, pagers, personal digital assistants (PDAs), web-phones, etc.

Of course, all the devices target different needs and users' types and, therefore, offer an equally variegated range of characteristics such as storage, display capabilities (such as screen size and color depth), wireless network connections, limited bandwidth, processing power and power consumption. These constraints currently pose several challenges on the designer since both the delivery and the presentation of complex personalized applications must take into account the limitations of the device.

*This research work was financially supported by the Italian FIRB 2001 project number RBNE01WEJT "WEB-MiNDS" (Wide-scalE, Broadband MIddleware for Network Distributed Services).

The key to meet the demands in this heterogeneous environment is the adaptation of contents to the capabilities of the devices and communication systems.

In this paper, we present content adaptation services for the Mobile Web as implemented on a programmable proxy infrastructure called Scalable Intermediary Software Infrastructure (SISI). In particular, we present the implementation on top of our SISI framework of Content Selection markup language DIWG, 2005 that was recently released (May 2nd, 2005) by the World Wide Web Consortium (W3C) Device Independence WG (DIWG) [http://www.w3.org/2001/di/].

2. PROXY FRAMEWORKS FOR MOBILE COMPUTING

One of the current research trend in distributed systems is how to extend the traditional client/server computational paradigm in order to allow the provision of *intelligent* and *advanced* services.

This computational paradigm introduces new actors within the WWW scene, the intermediaries Barrett and Maglio, 1999; Luotonen and Altis, 1994, i.e. software entities that act on the HTTP data flow exchanged between client and server by allowing content adaptation and other complex functionalities, such as geographical localization, group navigation and awareness for social navigation, translation services, adaptive compression and format transcoding, etc.

Several popular proxy systems, such as RabbIT [http://rabbit-proxy.sourceforge.net/], WebCleaner [http://webcleaner.sourceforge.net] and Privoxy [http://www.privoxy.org/] provide functionalities for text and image transcoding, removing cookie and GIF animations, removing advertisement, banner, Java Applets and JavaScripts code, protecting privacy and controlling access.

Web Based Intermediaries (WBI) [http://www.almaden.ibm.com/cs/wbi/] is a programmable proxy whose main goal is to simplify the development and the deployment of Web intermediary applications (i.e. applications that deal with HTTP information flows) and in particular content adaptation and personalization.

AT&T Mobile Network Rao et al., 2001 is a proxy-based platform designed to provide personalized services to users of mobile devices. It provides a modular architecture that allows an easy definition of new functionalities on top of a programmable proxy. The modularity is achieved through three different abstractions: devlets that provide protocol interfaces to different mobile devices; infolets that collect information from different data sources and applets that encapsulate the service's functionalities.

MARCH Ardon et al., 2003 is a distributed content adaptation architecture that provides functionalities for adapting Web content according to the capabilities of client devices that access to Internet services. An important feature

of this architecture is that it allows the dynamic composition of services for a given set of operating conditions (CPU, memory, etc).

3. SCALABLE INTERMEDIARY SOFTWARE INFRASTRUCTURE(SISI)

In this section we provide a brief description of the SISI architecture that we are currently developing and provide an overview of its implementation and a brief description of the mechanisms we used to personalize the services (for each user).

Our framework is based on top of existing open-source, mainstream applications, such as Apache Web server and mod_perl, because, first of all, of the quality of these products of open-source development and, secondly, because the results will be widely usable given the popularity of the Apache Web server.

SISI framework focuses on providing a simple approach for assembling and configuring complex and distributed applications from simple basic components. The idea is to provide functionalities that allow programmers to develop services without taking care of the details of the infrastructure that will host these services. Therefore, SISI provides a modular architecture that allows an easy definition of new functionalities implemented as building blocks in Perl. These building blocks produce transformation on the information stream as it flows through them. For a detailed description of SISI architecture and its implementation please refer to Grieco et al., 2005; Colajanni et al., 2005.

Users' profiles configuration. User and device profiling is a very important characteristic because of the heterogeneity of the devices and services of the Mobile Web. Users need multiple profiles according to their status or preferences, and, then, different configurations that must be easily switched to as well as modified. Of course, configuration must be performed by the user itself.

Our approach in managing user profiles is to explicit ask the user what s/he needs and use this information with a rule-based approach to personalize the content. In particular, users have to fill-out forms to create new profiles[1] and to modify or delete the existing ones.

Furthermore, because different client capabilities of the devices used to access the Web, users must be allowed to specify, in their device's profiles, the needed parameters both for devices and services. For example, when a user connects with a PDA s/he could want to be displayed only black and white im-

[1]The term profile, in this paper, is referred to SISI profile, i.e., the set of services that are configured (with their own parameters) by the end user. When referring to other kind of profiles, we will use the full definition, like CC/PP profiles, for instance, in Section 4.

ages or not given images at all in order to save bandwidth. To this aim the user only has to activate the corresponding profile previously saved into the system.

SISI allows a simple configuration mechanism for adding, removing, enabling or disabling functionalities and a more complete configuration mechanism where service parameters can be specified by asking the user to fill-out Web-based forms (for example, by providing a downgrade quality image for a transcoding service).

4. CONTENT ADAPTATION SERVICES

Adapting services to capabilities and preferences of different users and devices is a research issue addressed in both the mobile and the universal access research communities. The mobile community has mostly been concerned with device capabilities and technical problems, and has thus been more oriented toward solutions with a common presentation for all devices. Then, we first describe our implementation of the Working Draft DIWG, 2005 of the Device Independence Working Group of the W3C and next we show how SISI profiles interact with the CC/PP profiles.

Content Adaptation Independence Model. The W3C is currently working on activities relevant to content adaptation within the Device Independence Working Group. This Group is studying techniques for content adaptation, authoring and presentation techniques for Web content in order to match the different capabilities of client devices to which contents have to be delivered to.

In order to make the Web accessible anytime and anyhow, in particular by supporting other than many access mechanisms also many modes of use (including visual and auditory ones), they provide some important suggestions or principles as described in DIWG, 2003. Finally, this Working Group has produced documents for CC/PP recommendation, and documents that discuss authoring challenges and techniques DIWG, . From these documents it is possible to achieve information about authoring-related principles, and in particular out interest is mainly devoted to the need of, whenever possible, reusing content across multiple delivery context, adapting presentations independently from the access mechanisms.

Finally, a draft on Content Selection markup language DIWG, 2005 discuss how to express alternate contents or resources for their delivering to end users, by changing the presentation according to the capabilities of the requesting device.

Device Independence. The draft on Content Selection DIWG, 2005 specifies a syntax and a processing model for general purpose content selection or filtering. Selection involves conditional processing of an XML information ac-

cording to the results of the evaluation of expressions. Using this mechanism some parts of the information can be selected while other not delivered, automatically adapting the original content according to particular accessibility rules.

In the SISI Content Selection service (SISI CS service) we are implementing the specifications of the draft. We have currently implemented the conditional expressions, and the conditional expressions that return values, that can be further used to check if a particular piece of content is to be included for processing in the Web page delivered to end users (see Figs. 1 and 2). The remaining part of the specification is currently under testing or development.

If the end user aims to use the CS service, s/he has to configure service's parameters in order to meet the capabilities of the accessing terminal device. If the used device is a mobile one with a small display, the user can accordingly set height and width parameters, specifying the number of pixels that the device is able to support.

The CS service intercepts user's request, reads the user's profile and the service user-defined parameters, retrieves the Web page from the Web server (the original Web pages augmented with the `DISelect` markup expressions), performs some computation and, finally, delivers only the content that satisfy the specified conditional expressions and rules. Computations include the parsing of the HTML Web page to pull out the `DISelect` tags, invocation of the functions according the matched expressions to validate the values of the corresponding variables.

```
<br>
<sel:variable name="nColors" value="di-cssmq-monochrome()"/>
    <sel:if expr="$nColors">
        <p>Your device can display monochrome images.</p>
    </sel:if>
<center>
    <table cellspacing=30>
    <tr><td><img border=0 src="gouimg/gauguinmonochrome.jpg"></td></tr>
</center> ...
```

Figure 1. An example of conditional expressions.

The function `di-cssmq-color(0)`, in the example shown in Fig. 2, returns the number of colors supported by the accessing device. If the value is not zero, the paragraph is included in the delivered content by displaying the value of the `nColors` variable in place of the `sel:value` element. The presentation of the Web content will change accordingly to the value of `nColors`. For example, if the user uses a device that supports only color depth 256, then the CS service will retrieve only the embedded images with the specified resolution. Of course, the embedded images are available on the Web server at different color resolution, but only that one that matches the conditional expression is provided and displayed.

```
<br>
<sel:variable name="nColors" value="di-cssmq-color(0)"/>
    <sel:if expr="$nColors">
        <p>Your device can display <sel:value expr="$nColors"/> different colors.</p>
    </sel:if>
<center>
    <table cellspacing=30>
        <tr><td><img border=0 src="gouimg/gauguincolor.jpg"></td></tr>
...
```

Figure 2. An example of conditional expressions that return values.

Device Independence and CC/PP. The configuration of SISI CS service consists in specifying the parameters in order to meet the capabilities of the accessing terminal device. The W3C specifications of CC/PP (Composite Capabilities/Preferences Profiles) are used to specify the parameters. A CC/PP profile includes information on the device capabilities (e.g hardware and software characteristics, operating system, etc) and some user preferences.

In practice, a CC/PP profile contains a number of CC/PP attribute names and associated values that are used by a server to determine and deliver the most appropriate content to the client device. A set of CC/PP attribute names, permissible values and associated meanings constitute a CC/PP vocabulary. It is possible that metadata (user preferences and device capacity) are created in different time and resources can be in different place (i.e. vendor profile for a specific device).

To aggregate all the attributes there are the remote (or indirect) references: each reference links to a collection of default characteristics (specially useful for devices with limited bandwidth). This default characteristics are, then, integrated with the values specified by the user in order to realize the current profile and are delivered to the content server using an exchange protocol. The HTTP Extension Framework is a generic extension mechanism for HTTP/1.1 designed to interoperate with the existing HTTP applications and is the proposed CC/PP exchange protocol. The W3C Working Draft CC/PP Working Group, talks about the requirements, assumptions and goals of the CC/PP framework.

In our implementation, the set of rules to determine the server behavior on the value of the profile components is represented by the DISelect markup expressions that are evaluated accordingly to the DISelect variables populated by CC/PP profiles. Within SISI, the CC/PP profile can be used in different contexts: a remote CC/PP profile can fill in the values for the (CC/PP) attributes to be used by Content Selection service but can also be (locally) stored for configuration parameters of other services.

Currently there are not many implementations that can support the CC/PP protocol. Examples of implementations are DELI (Open Source Delivery Context Java Library supporting CC/PP and UAProf) Butler, , PANDA - Skunk (An Open source CC/PP test bed, plug-in for MS's IE and integration CC/PP with

P3P) Research, , and an implementation on Apache Web Server Papachristos and Markatos, using a proxy server.

SISI for the Mobile Web. From the architectural point of view, the programmability and users' profile management make SISI an interesting tool for the Mobile Web since it fits well in the proxy scenarios described in CC/PP Working Group, ; Papachristos and Markatos, .

In the first scenario, two proxies are used in the connection, a client proxy and a server proxy, to exchange CC/PP information and then to implement the protocol. SISI can easily manipulate request/response entities, and, therefore, can be placed either as client proxy or server proxy (or both) with two different modules (CC/PP aware) in the HTTP life-cycle.

The second scenario requires includes a server proxy and a client with browser CC/PP enabled (e.g., Panda Yasuda, ; Yasuda et al., 2001 from the Keio University) that implements transformations from HTTP to HTTP + CC/PP incorporating the client proxy functionality. SISI can be easily used as a standard HTTP proxy, to read the modified HTTP GET which carries the profile or the profile difference. The information extracted is used to retrieve the right content from the Web server or to adapt the response received to the access device capabilities. In case the profile is an inline reference to a profile repository, SISI can be easily programmed to cache profiles on behalf of the client.

The third approach is to implement a CC/PP aware Web server communicating with an intermediary client proxy. SISI can support the functionality of the client proxy, that is, to convert the standard HTTP headers sent by the Web client to extended HTTP headers that contain CC/PP headers (HTTP headers + CC/PP headers).

The last is the ideal approach where both server and client are capable of sending and receiving CC/PP, then, if we want use our framework, it should be a Web server CC/PP aware.

From the previous argument it turns out that SISI framework can be designed to support CC/PP and bridge the gap between the CC/PP-aware device and the server's delivery of device-appropriate content.

5. CONCLUSIONS

Without a full support toward the development of new services, advanced mechanisms cannot be easily realized, deployed and managed, while quick prototyping and manageability represent crucial requirements to assure that programmers can quickly respond to mutations of data format, content or standards that are so common in the Web, nowadays.

Therefore, our objective, in this paper, was to show how our intermediary and programmable framework can be used to quickly and efficiently allow the

deployment of services that can make navigation on the WWW from mobile terminals an enjoyable and non frustrating experience.

As a final consideration, we believe that SISI profiles are a key factor in providing an effective personal environment where each user can create/modify/delete profiles as well as easily switch among profiles as s/he needs. The integration of SISI profiles (that contains the proxy services activated for each HTTP request) with the CC/PP profiles is particularly fruitful, especially if seen in the context of the results presented in Bettini and Riboni, 2004 where an architecture for profiles and policy management is presented.

References

Ardon, S., Gunningberg, P., LandFeldt, B., Ismailov, Y., Portmann, M., and Seneviratne, A. (2003). MARCH: a distributed content adaptation architecture. *Intl Journal of Communication Systems, Special Issue: Wireless Access to the Global Internet: Mobile Radio Networks and Satellite Systems.*, 16(1).

Barrett, R. and Maglio, P. P. (1999). Intermediaries: An approach to manipulating information streams. *IBM Systems Journal*, 38(4):629–641.

Bettini, C. and Riboni, D. (2004). Profile aggregation and policy evalulation for adaptive internet service. In *Proc. of The First Annual Intl Conf. on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous 2004).*

Butler, Mark H. Open Source Delivery Context Java Library supporting CC/PP and UAProf. http://www.hpl.hp.com/personal/marbut/DeliUserGuideWEB.htm.

CC/PP Working Group. Composite Capabilities/Preferences Profiles: Requirement and Architecture. http://www.w3.org/TR/CCPP-ra/.

Colajanni, M., Grieco, R., Malandrino, D., Mazzoni, F., and Scarano, V. (2005). A scalable framework for the support of advanced edge services. In *Proc. of HPCC-05.*

DIWG. Authoring techniques for device independence. http://www.w3.org/TR/2004/NOTE-di-atdi-20040218/.

DIWG (2003). Device independence principles. http://www.w3.org/TR/di-princ/.

DIWG (2005). W3C Working Draft: Content Selection for Device Independence (DISelect) 1.0. http://www.w3.org/TR/2005/WD-cselection-20050502/.

Grieco, R., Malandrino, D., Mazzoni, F., Scarano, V., and Varriale, F. (2005). An Intermediary Software Infrastructure for Edge Services. In *Proc. of SIUMI'05.*

Luotonen, A. and Altis, K. (1994). World-Wide Web proxies. *Computer Networks and ISDN Systems*, 27(2):147–154.

Papachristos, C. and Markatos, E. A CC/PP aware Apache Web Server. http://dcs.ics.forth.gr/Activities/papers/ccpp_position.pdf.

Rao, C., Chen, Y., Chang, D., and Chen, M. (2001). imobile: A proxy-based platform for mobile services. In *Proc. of WMI 2001.* ACM Press.

Research, LSB. Research Project on Location Based Web Service. http://romeo.slab.sfc.keio.ac.jp/.

Yasuda, K. Implementation and Evaluation of Keio CC/PP Implementation. http://yax.tom.sfc.keio.ac.jp/panda/.

Yasuda, K., Asada, T., and Hagino, T. (2001). Effects and Performance of Content Negotiation Based on CC/PP. In *Proc. of The Second Intl Conf. on Mobile Data Management (MDM '01)*, pages 53–64, London, UK. Springer-Verlag.

CREATING AND PERFORMING SCENARIOS
FOR MOBILE SERVICES SUPPORTING MOBILE WORK IN EXPOSED PHYSICAL ENVIRONMENTS

Bente Skattør
AgderUniversity College, Faculty of Engineering and Science, Grooseveien 36, N-4876 Grimstad NORWAY

Abstract: Scenarios are widely used, explored and discussed in HCI. This paper addresses creating and performing scenarios in exposed physical environments, i.e. at construction sites. In this environment, little or no mobile ICT supports construction workers work. In an iterative design process with a user-centered approach, the project has created and explored ideas concerning actual use of mobile ICT supporting knowledge processes in exposed environment. The design process was based on three steps: workshops, shadowboxing (i.e. using dictaphones and mobile phones in real setting), and prototyping. By focusing and working with work incidents and related scenarios, we obtained the participation and involvement of the construction workers throughout the entire study. During shadowboxing, i.e. exploring by doing, the workers gained first-hand experiences and understanding (reflection-in-actions) of existing and future scenarios. This gave them a basis for mapping out when it was appropriate to use handhelds, and when it was not, and why.

Keywords: scenarios; mobile work; user-centered design; mobile services; reflection-in-action; knowledge processes, Contextual Design.

1. INTRODUCTION

Scenarios are frequently used when designing systems. There are already established approaches like scenario based design (Caroll 2000) and Contextual Design that support the use of information collected from observations to create scenarios (Beyer and Holtzblatt 1998). Often such approaches are used in further changes of existing systems, but they can also

be used in new application areas. One new, exciting area is the use of mobile ICT among mobile workers, or non-office workers, that normally do not depend on ICT to perform their work. Examples of such "non-office" workers are carpenters, electricians, installers, plumbers, ship surveyors, drivers, nurses and home-health carers. Many of the non-office workers have a work situation where they might be exposed to a hard physical work environment, e.g. rough climate, dust and noise.

The use of scenarios related to mobile services is an emerging field that has been applied in various settings. Bødker (1999) uses video to study the work at a combined district heating and power plant. She uses scenarios to support tensions between reflection and actions, between typical and critical, and between plus and minus situations. Binder (1999) suggests videotaping as a way for the electricians in an industrial setting to improvise scenarios with their own design of low-fi artefacts. Brandt and Grunnet (2000) use drama while they explore settings, scenarios and props in their studies of refrigeration technicians in supermarkets and families in their home. Iacucci and Kuutti (2002) create sessions using everyday life as a stage and the opportunity for participants to exercise reflection-in-action. Svanæs and Seland (2004) report workshops where end users were enabled to design mobile systems through scenario building, role playing and low-fidelity prototyping. Some of their workshops were related to nurses in hospitals. Garabet et al. (2002) elaborate on how performance art can be used to elicit information about device design and how art(ifacts) were used to spark behaviour and debate. Newcomb et al. (2003) use scenarios during design and usability testing of a mobile shopping application on a PDA in a shopping environment. Messeter et al. (2004) discuss the use of scenarios regarding implications for the design of mobile IT devices, focusing on coping with multiple social contexts and configuration and connectivity of mobile services.

There are, therefore, several examples of research related to scenarios in different mobile settings. Nonetheless, further research into, and the use of, scenarios in mobile settings are required to meet the challenges posed by this emerging field of mobile technology.

Even though blue-collar workers were some of the first to adopt mobile ICT (Churchill and Munro 2001) there has been a lack of research on blue-collar workers and design and use of mobile ICT (Brodie and Perry 2001). Some exceptions exist and probably the best known is the study by Luff and Heath (1998). This study might be an example of the importance of identifying mobile services that are appropriate to supporting mobile work. They discovered that the system hindered mobile collaboration because the wrong features of mobile work were chosen to be supported (Luff and Heath 1998).

To obtain participation and involvement among the construction workers this research has a user-centred approach. The exploration of their work and its constraints at the construction sites is grounded in an ethnographical approach with use of Contextual Design (Beyer and Holtzblatt 1998). Through the entire iterative design process we have focused on work incidents while creating and performing scenarios. The iterative process was based on workshops, shadowboxing (reflection-in-action) and prototyping.

This paper is organised as follows: Section 2 presents the research design including the research aims and the iterative design process. Section 3 reports on the results. Section 4 evaluates the design process by discussing the scenarios, and section 5 concludes the paper.

2. RESEARCH DESIGN

This research reports on experiences from an ongoing development and research project within industrialized residential construction in Norway. Within this type of manufacturing, ICT support project management and the design process of construction. However, little or no ICT supports the work carried out at the building sites. The physical work environment at the sites is characterised by variation in temperature, humidity, light, noise, dust and dirt. Today approximately 11% of workers in Norway work as craftsmen (Statistics Norway 2004) and workers within industrialized residential construction are part of this group.

This development and research project is in an early design phase and aims to develop mobile ICT to support the following three areas: 1) communication between the project team at the offices and the operation team at the building site, 2) knowledge processes at the building site, and 3) logistics at the building site. This paper reports experiences when designing mobile ICT supporting knowledge processes. The aims of this research were:

- To create ideas for the actual use of mobile ICT supporting knowledge processes like retrieval, storing, transferring and re-use.
- To test out and explore some of the emerging ideas and scenarios in their exposed physical environment.
- To design a prototype to support some of the emerging ideas for possible mobile services.

2.1 An Iterative Design Process

Knowing that the construction workers at the sites do not use mobile technology to support their work and acknowledging that they are the domain experts, this research adopted a user-centred approach in order to

obtain participation and involvement among the workers. The exploration of their work and its constraints at the construction sites is grounded in an ethnographical approach using Contextual Design (Beyer and Holtzblatt 1998). In order to reach the aims in the project and since we were going to study a novel field, we conducted an iterative design process based on *workshops*, *shadowboxing* (reflection-in-action) and *prototyping*. Through all the steps and iterations, we focused on work *incidents* and created and performed related scenarios. Figure 1 illustrates the iterative design process.

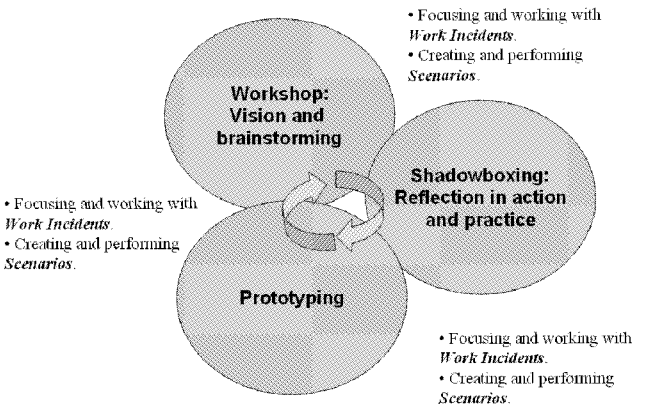


Figure 1. The iterative design process in this research project

The study was carried out over a period of six months by four full-time participants, three construction workers and one researcher. A mixture of qualitative techniques like observations, interviews, paper reviews (work process documents, checklists, and similar), taking pictures, video and recording voice was used. About twenty construction workers at the construction site were observed during work. Time spent with these workers varied from a few hours up to several days. The workers included professionals like carpenters, concrete workers, electricians, plumbers, bricklayers, painters, digger drivers and crane drivers. Twelve of these workers were also interviewed in depth. Furthermore, eight workers at the construction hut and the headquarters were interviewed. These workers cover occupations like gang leaders, project leaders, working managers, and architects.

Before describing the main steps in the iterative design process, the work incidents will be explained and exemplified.

Work incidents related to knowledge processes

In order to support knowledge processes with ICT it is important to structure the different knowledge types and the knowledge processes (Alavi and Leidner 2001; Fagrell et al. 1999; Kucza 2001). The preparatory work for this research was grounded in the work of Alavi and Leidner (2001). They report and discuss knowledge types like tacit knowledge (Polanyi 1966), explicit knowledge, individual knowledge, social knowledge, know-how, know-about, know-why, know-with and know-when (Alavi and Leidner 2001) and furthermore, they report on knowledge processes like retrieval, storing, transferring and re-use.In this study we decided to look for *work incidents* or situations that could exemplify knowledge processes at the construction sites. That was *incidents* that required some kind of action in order to proceed or to be solved. A *work incident* could be an experience, a quality problem, a question, a suggestion for improvement, a request, or an order. Possible *work incidents* could be related to normal working procedures or routines, related to complex activities or complex problems, related to lack of information and lack of quality of work. The knowledge processes were be examined with respect to different levels like individuals, teams and construction projects. Table 1 presents two examples of such *work incidents*.

Table 1. Examples of work incidents

Example no.1: Changing position of lay shafts in bathrooms. One or two apartments are often completed at an early stage in order to be used as showrooms for potential customers. This was also done in the case project. During assembly work in the bathroom in this apartment, some weak points in fasteners for the mirror were uncovered. At first, the assembly work was done according to the drawings. However, based on experiences gained while making minor adjustments to the mirror, the position of several of the lay shafts had to be changed. This experience or correction had to be registered and transferred to the more than 300 other bathrooms to be built.
Example no.2: Bad quality of pillars. Early in the project a participant discovered a lot of the pillars that arrived at the construction site were of poor quality. Many of the pillars were crooked, probably because they were made from rapid-growing trees. The carpenters spent a lot of time on grading and piling up the pillars. Between 40-70% of the pillars could be used. During the project a new type of pillars was tested out. These pillars were made of glued wood, and were then used in every case. Even though the new pillars were more expensive than the old type, money was saved because of less work.

Workshops

On a weekly basis we had workshops at the construction hut lasting about 2 hours. At these workshops we had brainstorming sessions, worked with

our vision, discussed work incidents and other findings, and created and discussed scenarios using work incidents. One of the main steps in Contextual Design (Beyer and Holtzblatt 1998) is to envisage how new technology will address the user work practice by creating a high-level story (vision) of how the work will be changed and working out the details of this story at the level of the task. During the brainstorming sessions and while working with the vision, we used mock-ups as suggested in Contextual Design (Beyer and Holtzblatt 1998).

Shadowboxing: scenarios of reflection in action and practice

In order to create ideas and achieve insight into how, when and why mobile ICT can support work in mobile settings, the carpenters started to use dictaphones[1] as props during work. They had to pretend that the dictaphone was a real mobile application, i.e. perform future use of mobile services. Each time a *work incident* arose, they had to register it with the 'mobile' application, that is, make an audio-record. This exercise we called "Shadowboxing". After 6-7 weeks they started to use mobile phones (SonyEricsson P900) as props. With these mobile phones the workers took pictures and made audio-records and videos of the *work incidents*. This approach has strong similarities to what Buchenau et al. (2000) call 'Experience Prototyping'. The users 'experience it themselves', rather than witnessing a demonstration of someone else's experience. Furthermore, this approach stresses the importance of 'being there' for the understanding of the contextual factors, as Oulasvirta et al. (2003) do in their case studies in bodystorming, that is brainstorming conducted 'in wild'.

As with Bødker (1999), this research finds it important to create trial use situations as part of the design process, in order to stage users' hands-on experiences with the future. Also it is important to investigate ways in which users can feed back reflections over work and trial use that are directly anchored in the specifics of use in work situations to designers (Bødker 1999).

Since construction workers do a physical work and are continuously using their hands, it is of interest how and when they use the dictaphone or the mobile phone to record their work incidents. We were challenged to identify appropriate times of interactions where the mobile application could offer assistance to the user while working. By exploring by doing, the hope

[1] The dictaphones have different types of size and shape. Also, they are different as to how functions are used (differences in menus) and different in functionality richness. Three types of dictaphones (Apacer Audio Steno BP300 MP3 Player 256MB, MP3 Spiller - Asono Beatsound 256MB, MP3 Spiller - Asono 50f 256 MB) were used.

was that this could provide first-hand understanding and experience of existing and future scenarios.

Figure 2. Tom uses the P900 to record a work incident.

Prototyping

According to Contextual Design we used User Environment Design and paper prototyping (Beyer and Holtzblatt 1998). In order to visualise our ideas we made a prototype for handling the work incidents, e.g. registering, updating, follow-up and reporting. According to our vision the prototype covered mobile services for handhelds at the construction site and an application for stationary PCs at a construction hut or an office. The prototyping was done in three iterations.

3. RESULTS

3.1 The Vital Importance of Work Incidents

The work incidents were mainly discovered by the participants during work or by observing other construction workers. It became evident that the *work incidents* were of vital importance.

Documenting and working with the *work incidents* gave us deep insight into situations and problems at the construction site. Also, trying to solve the

work incidents generated scenarios in a naturally way. The group created scenarios by thinking of possible new solutions while discussing situations like: How did the incident arise? How could it be communicated and registered? How could it be solved, and who could participate in solving it? What could we learn from the incident? Moreover, talking with other workers that were not members of the project about *work incidents* gave further insight and inspiration for new scenarios.

Working with the work incidents and their related scenarios certainly contributed to the engagement and involvement of the members through the whole project. This was of vital importance and came to light every time we solved a *work incident* in the project. Through this, the members felt that they were contributing with improvements both on the spot and for the future. Furthermore, since the *work incidents* were based on their personal experience, the creation of scenarios became more vivid and engaging.

The first two workshops when we brainstormed on a possible vision, the carpenters did not contribute a lot. However, as soon as we started to work and solve *work incidents* and tried to 'place' these into the vision, this changed. Now they were the experts and were overwhelming with ideas and suggestions. During this process the users invented and told stories. Furthermore, as we started to use mock-ups with simple illustrations, i.e. in this case stick figures men, they became engaged in the process. The workshops revealed open-ended (Bødker 1999) and wide-ranging scenarios on a conceptual level. However, as the process proceeded, the scenarios became more and more detailed. It appears that this was connected with the use of the *work incidents*. The further we went into detail in order to solve the work incidents, the more detailed the scenarios became.

Fore each of the three iterations of prototyping we used a number of the *work incidents* and their related scenarios as test cases for the prototype. This was very useful and made us focus on concrete and real issues, problems and processes. Furthermore, this made it possible to validate whether the prototype was able to handle the workers daily *work incidents*.

3.2 Shadowboxing was a Very Fruitful Activity

We have found that shadowboxing, i.e. using a dictaphone or a mobile phone as props in order to register the work incidents, was a very fruitful activity. Setting the stage for 'context of practices' and 'reflection-in-action' (Schön 1983) enforced ideas about possible use. And, by exploring by doing, the users gained first-hand experiences and understanding of existing and future scenarios. This gave the workers a basis for mapping out when it was appropriate to use handhelds, and when it was not, and why. A great many

experiences were revealed during this activity. Table 2 contains three of these.

Table 2. Examples of when it was appropriate to use handhelds, when it was not, and why

Example 1: In general, the workers found the use of handhelds was strongly related to what kind of work incident they were handling. For instance, very seldom it was useful to register the work incident at a very early stage when it was related to experiences or knowledge processes. Frequently, they had to do some trial-and-error work before they actually could register or describe the work incident. However, when the work incident was related to logistics, they saw the possibility to reduce the time used in searching for materials. Instead of going around the site to look or ask other workers, they could use the handheld for locating them. Based on the location given by the handheld, they then could go and fetch the materials, or order the materials directly by using the handheld if the materials were out of stock.
Example 2: They learned from experience that the use of pictures or video from a mobile phone was more suitable than the audio-record when the work incident was complex or detailed. They found it easier to explain the situation with the use of a picture.
Example 3: Many of the work procedures in residential construction are well-defined through certain sequential building processes. For instance, first the concrete has to set, then come the ground beams, then the pillar for the walls, then the electricity work, then the plumbing, then the insulation procedure, then the planking or plaster, then the painting and so on. But since many types of professionals and workers are involved at different times for different periods, the management of planning, informing, coordinating, and following up progress is a complex and challenging activity. Today, the gangers have the operational management for this out at the sites. Since the sites are often large and hard have an overview of, the gangers cannot easily reach the workers, and visa-versa. Consequently, the participants see possibilities to support the building progress. They can use a handheld to report on progress, progress problems, or even better, before they get into problems.

Since the study lasted for months, each participant was alone most of the time while shadowboxing at the construction site. This gave them time to consider and evaluate their own thoughts and ideas, what Schön (1993) calls reflection-in-action, i.e. thinking about what we are doing. However, when two or more were together, the shadowboxing was expressed verbally and intertwined with discussion. This exercise of exploration and evaluation provided confirmation or rejection of the scenarios based on real experiences.

Kuutti et al. recognise a successful creative process in the performance where there are at least two things happening. Firstly, there is performing and interaction with the physical reality, by which participants take action in the physical world and change it during symbolizing activities (Kuutti et al. 2002). In this study we find this to be fulfilled during shadowboxing. The dictaphones and the mobile phones were used in their real settings and symbolized new activities, or even in all probability, they have some of the mobile services needed at the construction sites (i.e. audio-recording, video, picture and registration of texts). Secondly, the participants interpret symbolizing actions in the changed environment, and aim at a shared

interpretation in a collaborative way (Kuutti et al. 2002). This was achieved when two or more were together at the site performing shadowboxing and expressing scenarios verbally. These interpreting scenes were intertwined with discussions, confirmations or rejections.

3.3 Mobile Technology in Exposed Environment

During the project the participants experienced many other aspects that might be of relevance for designing mobile technology in this kind of environment. Table 3 lists three such aspects.

Table 3. Aspects relevant for designing mobile technology in exposed environments.

Aspects	Description and comments
Handhelds as work-wear	Due to the natural setting at the construction site, the handhelds must withstand variation in temperature, humidity, light, noise, dust and dirt. In addition they must withstand tough usage by the workers, and to some degree be shock-proof. The handhelds have to be wearable, or more precisely, they have to become a part of the work-wear. Even though the project was not carried out during winter months, neither the dictaphones nor the mobile phone tolerated this usage and environment. None of them were robust enough.
Noise disturbance	**Example 1:** Because of noise the workers had a tendency to raise their voice while doing audio-recording. The consequence was bad audio-quality. **Example 2:** During video-recording with the mobile phone, some technical problems occurred. The participants tried to fix this, without success. To solve the problem they had to look for a quiet place. They could not concentrate enough to solve the problem in the noisy environment.
Culture	**Example 1:** In the beginning the construction workers had problems with audio-recording at the construction site. They complained about the poor sound quality of the dictaphones, and they had to record at home or in the car. However, based on further investigation, it came to light that this was not the case. In fact, they really felt uncomfortable when making audio-records at the site. They did not want other workers to see them in that situation, even when they were alone on a section at the construction site. However, there was a breaking-in period and after five-six weeks, they got used to using the dictaphones on site. **Example 2:** Even though the participants got used to the dictaphones, i.e. using it during their daily work, they still felt uncomfortable using them or the mobile phones in a particular setting. This became very evident when participants visited other places at the site to register an 'old' work incident, for instance taking a video instead of an audio-record, while some other workers not familiar with the work incident were present. The participants then felt like total strangers. This feeling in this particular setting lasted for the rest of the project.

Table 3 illustrates the importance of social and cultural aspects in the design and use of mobile technology. We fully agree with Visciola (2003) when she states that social issues are one of the important emerging research issues in research on wireless applications.

3.4 Aspects of the Scenarios

In the following important aspects during the exploration of the creating and performing of scenarios are highlighted:

Open-ended scenarios: In this research we have tried to facilitate open-ended scenarios (Bødker 1999) in order to give broad and conceptual answers. We found this to be the case in the beginning of the process when we worked with the vision and new ideas of mobile services. As the project proceeded, the users tended to work with more detailed scenarios related to *work incidents*, e.g. *closed* scenarios (Bødker).

Memorable and inspiring scenarios: According to the study of Oulasvirta (2003) bodystorming sessions make scenarios both memorable and inspiring. This was also the case in this research. The shadowboxing (i.e. bodystorming) with real work incidents and their related scenarios was thought to be inspiring by the participants. Furthermore, during the workshops' brainstorming sessions of possible scenarios based on the shadowboxing, it became clear that the work incidents and scenarios were felt to be memorable.

Roles of performance: Iacucci et al. (2002) emphasize three roles of performance in the design of interactive systems: exploring, communication, and testing. This research contains examples of all these roles. The exploring was performed while shadowboxing on the site. Further, the *work incidents* and their related scenarios were communicated in the project between all participants in all steps. And finally, the testing of the scenarios and our vision/concept was done during the prototyping iterations.

3.5 Problematic areas

Knowledge processes: One of the aims of the research was to create ideas for the actual use of mobile ICT supporting *knowledge processes*. Very early in the project we experienced that it was very hard to gather only *work incidents* related to knowledge processes. Most of the incidents were not an 'obvious case'. Frequently we spent a lot of time discussing whether the *work incidents* were within our scope, or not. So, after a while, we decided to register all *work incidents* that we discovered.

We ended up with 55 highly diverse work incidents. In order to separate *work incidents* related to knowledge processes from the other processes, further analysis and categorisations were necessary. The documentation was extended to include categories like seriousness, frequency, different classifications of causes, classification of production area, professions, and costs. Through this comprehensive analysis and categorising of the *incidents*, we found that about 30% were related to suggestions for improving the

building processes that were based on work experiences, about 26 % were related to quality problems (often cased by lack of knowledge or information), and about 17 % were related to communication between the projecting team and the operation team. The remaining, were related to other categories like logistics, project management and health and safety.

As we have gained a deeper insight into *work incidents* and scenarios related to knowledge processes, we are aware of much that we do not yet understand. Further research has to be done in order to understand and structure knowledge processes and knowledge types. However, we have gathered an overwhelming amount of complex information. All the *work incidents* are stored in the database of the prototype, including videos, pictures and voice. Thus we can see the beginning of a database containing information about knowledge and experiences. Also, we see possibilities of mobile services supporting discontinuities (i.e. *work incidents*), thus bridging the gap between intention and action (Gershman et al. 1999).

Documentation: During the discussions of scenarios at the workshops and the shadowboxing we were not able to document all the scenarios. We realise that we should have used audio recording or video during these discussions. Moreover, much of the process of documentation, especially the diary/log of the work incidents was very time consuming.

4. DISCUSSION

Despite the difficulties related to knowledge processes, we venture to say that the iterative design process was very helpful in order to achieve the research aims. The project has created and explored ideas relating to the actual use of mobile ICT to support knowledge processes in workers' exposed physical environment. Also a prototype was designed.

This study has illustrated the vital importance of real *work incidents* and the creating and performing of related scenarios. Hence, it could be of relevance to evaluate the iterative design process by evaluating and discussing the scenarios in this study. Just as Svanæs and Seland (2004) evaluate their workshops with respect to objectivity, reliability, validity and transferability we also evaluate the scenarios in the same way:

Objectivity: To what extent do the scenarios and the ideas originate from the users, and not from the facilitators or developers? In this project all the work incidents were discovered by the construction workers or through observation of them. At the beginning of the project the researcher had to help the users in how to create scenarios. However, after a few sessions, the construction workers were able to create the scenarios themselves. They were the experts and they knew it.

Reliability: Are the scenarios accurate in their description of the situations being studied? Since the scenarios most often emerged from real work incidents they certainly are in-situ and are highly relevant to the situation being studied. The scenarios created during the workshops were further elaborated and verified during shadowboxing, i.e. grounded in real settings. At the end of the project a cost analysis of the work incidents was done which revealed that most of the work incidents were costly to handle or solve. This supports the facts that the work incidents are relevant and important.

Validity (internal): Are the scenarios describing the important aspects of the situations related to the purpose of the research? The scenarios describe the actual use of mobile ICT supporting knowledge processes at the construction sites. However, we had to extend the scope to other types of processes and work incidents. The research also gave insight in such a way that the workers could map out when it was appropriate to use handhelds, when it was not, and why. Furthermore, we tested the validity of the prototype by using work incidents and scenarios, i.e. we visualised whether the ideas and design could handle the work incidents.

Transferability: Are the scenarios typical of the situations being studied, i.e. can the conclusion drawn from analysing the scenarios be generalized? Shortly after the fieldwork was finished, its results and experiences were presented to a group of managers in a large global construction company. This company builds (large) buildings, residences, roads, bridges, tunnels and so on. They found it both relevant and interesting and are now a partner in the main development and research project. This could indicate that the scenarios are of relevance for other construction sectors.

Another, and maybe more interesting, question is whether the scenarios could be generalized to other kinds of mobile work in exposed physical environments. Some of the conceptual scenarios might be, e.g. a handheld that makes it possible to register *work incidents* and transfer information into a database, and the handling of the *work incidents*. The scenarios that are detailed we find not appropriate to generalize. In all probability they are too specific for the construction sector, and some even too specific for industrialized residential construction. However, we see transferability of the scenarios within the different professions at the construction site, like carpenters, concrete workers, electricians, plumbers, bricklayers, painters, digger drivers and crane drivers.

However, what we find appropriate to generalize to other kinds of mobile work in exposed physical environments is the steps in this user centered design process to create the scenarios. Especially using *work incidents* and their related scenarios during shadowboxing might be relevant in other industry or sectors.

5. CONCLUSION

This research has resulted in a diversity of experience through creating and performing scenarios in mobile work in exposed physical environments, i.e. industrialized residential construction. In an iterative design process based on workshops, shadowboxing (i.e. using dictaphones and mobile phones as promps at the site) and prototyping, the project has created and explored ideas for the actual use of mobile ICT supporting knowledge processes. This paper has described shadowboxing as a very fruitful activity. Through exploring by doing, the users gained first-hand experience and understanding (reflection-in-action) of existing and future scenarios. This gave the workers a basis for mapping out when it was appropriate to use handhelds, when it was not, and why.

Moreover the research has successfully promoted the participation of users in an innovative design process. By focussing on work incidents, elaborating and working with the incidents at all the stages, we obtained participation and involvement among the construction workers through the entire design process. The use of work incidents and related scenarios was of vital importance. This paper has highlighted and discussed important aspects of the scenarios.

ACKNOWLEDGMENTS

I would like to thank all the workers participants for generously giving their time and effort during the project. Specifically, I wish to thank and credit Tom Ø. Hansen and Kris Sigurjonsson whose attitudes and inventions have inspired the fieldwork through the whole period.

References

Alavi, Maryam, and Dorothy E. Leidner. 2001. Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. MIS Quarterly 25 (1):107-136.

Beyer , Huge and Karen Holtzblatt 1998. Contextual Design - Defining Customer-Centered Systems, Morgan Kaufman Publishers, ISBN: 1558604111

Binder, T. (1999) Setting the Stage for Improvised Video Scenarios, ext. Abstracts CHI'99, (pp230-231)

Brandt, E., Grunnet, C. (2000) Evoking the future: drama and props in user centred design, In: Cherkasky, T., Greenbaum, J., Mambrey, P. (Editors), Proceedings Participatory Design Conference, New York, CPSR, 11– 20

Buchenau, M., Fulton Suri, J. (2000) Experience prototyping. Proceedings DIS2000, Designing Interactive Systems, New York City, USA, ACM Press, 424–433

Bødker, S. (1999) Scenarios–setting the stage for reflection and action in user-centered design. Proceedings Hawaii International Conference on System Science.

Carroll, J. (2000) Making Use: Scenario-Based Design of Human-Computer Interactions. MIT Press

Fagrell, Henrik, Fredrik Ljungberg, Magnus Bergquist, and Steinar Kristoffersen. 1999. Exploring Support for Knowledge Management in Mobile Work. Paper read at The European Conference on Computer Supported Cooperative Work (ESCW'99), at Copenhagen, Denmark.

Iacucci, G., Iacucci, C., Kuutti, K. (2002) Imagining and experiencing in design, the role of performances Proceedings of the second Nordic conference on Human-computer interaction, October 2002, 167-176

Iacucci, G. Kuutti, K. (2002) Everyday Life as a Stage in Creating and Performing Scenarios for Wireless Devices. Personal and Ubiquitous Computing, 6:299-306.

Garabet, A., Mann, S., Fung, J. (2002) Exploring Design through Wearable Computing Art(ifacts). Extended Abstracts CHI 2002, Minneapolis, Minnesota, ACM Press, 634-635

Gershman, Anatole V., Joseph F. McCarthy, and Andrew E. Fano (1999). Situated computing: Bridging the gap between intention and action. Paper read at The Third International Symposium on Wearable Computers, at San Francisco, USA.

Kucza, T. 2001. Knowledge Management Process Model: Technical Research Centre of Finland,VTT Publications.

Kuutti, K., Iacucci, G., Iacucci, C. (2002) Acting to Know: Improving Creativity in the Design of Mobile Services by Using Performance. Proceedings on conference Creativity & cognition, 2002 October 14-16, ACM Press, 95-102

Luff, P., Heath, C. (1998). Mobility in Collaboration. Paper read at Proceedings of the 1998 ACM conference on Computer supported cooperative work, at Seattle, Washington, United States.

Newcomb, E., Pashley, T., Stasko, J. (2003) Mobile Computing in the Retail Arena, Proc. CHI 2003, April 5-10, 2003, Ft. Luaderdale, Florida, USA.

Messeter, J., Brandt, E., Hasle, J., Johansson, M. (2004) Contextualizing Mobile IT. DIS2004, August 1-4, 2004, Cambridge. Massechusetts, USA. 27-36.

Oulasvirta, A., Kurvinen, E., Kankainen, T. (2003) Understanding contexts by being there: Case studies in bodystorming. Personal Ubiquit Computing Vol. 7, 125-134.

Polanyi, Michael. 1966. The Tacit Dimension. London, UK: Routledge and Keoan.

Schön, D. (1983) The Reflective Practitioner. How Professionals Think in Action. MPG Books Ltd, Great Britain, ISBN1857423194.

Statistics Norway (2004) Employment divided in professions and gender. Yearly average 2002-2004. http://www.ssb.no/emner/06/01/yrkeaku/tab-2005-02-02-02.html

Svanæs, Dag., Seland, Gry. (2004) Putting the Users Center Stage: Role Playing and Low-fi Prototyping Enable End Users to Design Mobile Systems. CHI 2004, April 24-29, Vienna, Austria (pp 479-486)

Visciola M. (2003) "Reflections on the User Centered Design Perspective in Rearch on Wireless Applications. Ubiquity, An ACM IT Magazine and Forum, Volume 4, Issue 8, April 16-22, 2003

HOTDESKING:
A Potential Link in the eWorker's Information Chain

Crystal Fulton
University College Dublin, Belfield, Dublin 4, Ireland, Crystal.Fulton@ucd.ie

Abstract: One of the major challenges of the flexible workplace is sustaining workflows while enabling mobile work. Hotdesking is intended to facilitate work in temporary workspaces in a mobile work environment. This study explored trends in information behaviour supporting work tasks through hotdesking in Canada and Ireland. Hotdeskers participated in semi-structured interviews about their information behaviour. Hotdeskers in both countries similarly identified access to electronic resources, reorganization of information, storage, permanent workspace, and mobile technologies as key items for successful hotdesking. The findings point to a particular information seeking behaviour among hotdeskers and suggest areas for future development of hotdesking arrangements.

Keywords: eWork, Hotdesking, Information Behaviour, Mobile work, Mobility

1. INTRODUCTION:

Mobile work can take place in a variety of contexts, including working from home, on the road, in hotels, in airport lounges, etc. Regardless of how work is located, however, a major challenge remains sustaining workflows through information access. The hotdesk model of working, traditionally a work environment designated for temporary use by multiple persons, is reputed to offer the eWorker the opportunity to reconnect with the organization, while simultaneously allowing the organization to reduce space requirements for permanent offices. By exploring the hotdesking work environment, we can increase our knowledge of the role hotdesking plays in the creation and adoption of information structures to facilitate information

flows. The object of this study was to increase our understanding of the information world of employees working in flexible work arrangements which include hotdesking. An examination of Canadian and Irish experiences with hotdesking was arranged through a government research partnership to facilitate research between the two countries and offers an important point of comparison between well-established and newer hotdesking arrangements.

2. HOTDESKING AND EWORK:

Alternative methods of working are features of work in both Canada and Ireland. While up-to-date information on the numbers of people working in alternative work environments is not always available, some organizations have attempted to count people engaged in different forms of work, and figures are available for the recent past. The Canadian Telework Association estimates that the number of eWorkers is approximately 1.5 million (CTA, 2001).[1] *Telework in Europe* (2000) identified one million or 7.1 per cent of working Canadians as eWorkers and ranked Canada as 6th worldwide for penetration of eWork in the workforce. The same report ranked Ireland 7th overall for implementation of eWork, with 40,000 eWorkers or 2.9 per cent of the working population involved in eWork. The Electronic Commerce and Telework Trends (EcaTT, 2000) study found that a higher number, 4.4 per cent, of the Irish workforce eWorks. Although Ireland has not experienced the same degree of implementation of eWorking arrangements as Canada, there is a movement toward increased adoption in Ireland, championed by key public officials (e.g., Noel Treacy, Minister for Science, Technology and Commerce, speech delivered at TWI, 2001). Hotdesking is seen as one means of supporting eWork initiatives in Ireland (TWI, 2001).

Hotdesking, or hoteling as it is sometimes called, provides a system of working in a shared work environment. Characteristically, the employee can access the equipment and technology necessary to perform work tasks or may plug in a laptop to access the organization's resources, but no one person has ownership of a given workstation. However, hotdesking may also be configured to allow working in temporary spaces in home, client, and organizational contexts. Through hotdesking, eWorking employees have the flexibility to work at a remote location for a period of time, returning to the organizational workplace to perform particular work functions, such as attending meetings, gathering information, etc.

[1] Figures on the numbers of eWorkers are difficult to estimate and information is not always gathered at regular intervals through formal channels. The numbers cited in this paper reflect the most current information available at the time of writing this article.

The hotdesking model may offer a valuable gateway to information for the mobile worker. The question of information exchange, both among eWorkers and between eWorkers and other organizational members, is seldom addressed in studies of eWork. Indeed, eWork still represents a relatively new phenomenon in work arrangements and little is known about the effect of this alternate work organization on information flows. However, Fulton (2000) revealed that one of the problems associated with remote work done from home has been isolation from resources, including colleagues and information. This study showed that information professionals working at home were often unable to complete work tasks because they lacked information. Since work tasks require access to information which often must be timely, changes in information access, as well as in information organization and storage, are extremely important to the work done by eWorkers. Hotdesking has the potential to provide a critical link in their information seeking, connecting eWorkers to information they are missing or cannot access in other work environments.

Several theoretical models of information seeking behaviour reveal patterns of information access and use that are relevant to the workplace, including the processes by which people recognize and resolve information needs. Wilson (1999) has shown that a person in a particular context who engages in information seeking behaviour may be challenged by such intervening variables as role-related or environmental factors. Cheuk (1998) notes that workers proceed through seven information seeking and using situations in the workplace: the beginning of a new task, the recognition of the need for understanding about the task, the development of ideas to execute the task, verification of ideas developed, problems with conflicting information, finalization of ideas, and the exchange of ideas with others. Leckie et al. (1996) also emphasize the central role of work tasks in the information seeking behaviour of professionals, noting that information needs arise from work tasks. In the case of eWork, employees who work in more than one location often alter their information seeking strategies to accommodate the movement of work to a different context (Fulton, 2000).

Approach:

This study explored the following question: *What role does hotdesking play in the information chain of eWorkers?* This research question raises other important issues. For example, what particular resources do employees use in the hotdesking environment, which they cannot access in their remote workplaces? How does hotdesking support the information seeking of eWorkers?

Data were gathered through semi-structured interviews and participant observation in multiple hotdesking cases to build a detailed picture of the information world of the eWorker in a hotdesking environment. Interviews

were organized around the Critical Incident Technique (Flanagan, 1954), a method of gathering important facts concerning behaviour by focusing on a defined situation, used in this case to explore participants' information behaviour while hotdesking. Changing patterns in information exchange were explored by asking participants to review, step-by-step, their use of information in their work.

A total of sixteen information professionals, divided equally among three different, large organizations and between Canadian and Irish workplaces, participated in this qualitative study. The number of participants in the study falls within the guidelines provided by qualitative research experts, such as Miles and Huberman (1994). Most often, participants were male (70% male; 30% female), were in mid-career, and had achieved high levels of formal education. All participants were employed in some form of information work spanning various organizational levels. Two of the organizations had established hotdesking practices; the third organization was experimenting with hotdesking and had completed a hotdesking pilot study. In the majority of cases, participants used hotdesking in combination with other forms of working, including home-based and mobile eWork.

Information Behaviour in the Hotdesking Environment:

Participants identified an array of resources as critical sources of information, including email, the Internet, company Intranets, and electronic databases. Participants also valued verbal contact with colleagues and clients, as well as certain print resources, such as business reports available only in printed format; however, electronic information held particular importance to participants as information that could be transported easily from workspace to workspace.

Although hotdesking was intended to facilitate information access, participants frequently identified situations where they were missing information, particularly electronic information. Remote electronic access was not always possible, nor easy; for instance, participants might not have direct access to electronic company files when working remotely. A common strategy for overcoming gaps in information was to keep multiple copies of documents in paper and electronic formats. However, creating multiple copies of information items also necessitated keeping track of one's most recently updated files.

Participants reported that they missed regular verbal interactions with colleagues, especially when colleagues were also hotdeskers. To fill in information gaps, participants relied upon their colleagues, and in particular, colleagues with permanent workspaces within the organizational office; however, hotdeskers were conscious of interrupting their colleagues, both when hotdesking at the central office and when working remotely, and reserved this option for more information critical situations.

Participants further attempted to cope by reorganizing their information environment. In particular, hotdeskers moved away from paper-based information sources whenever possible, since paper is heavy to carry and requires storage space. Laptop computers were especially important for information storage; hotdeskers relied upon their laptop computers to facilitate their movement from workspace to workspace within and outside the organizational office. However, only participant actually managed to implement a paperless office. Participants generally maintained printed files that were critical to current and ongoing work processes. These printed sources were stored for convenient access, in some cases in general office archives and at other times in assigned lockers and small filing cabinets. It was also common for hotdeskers to turn to their colleagues for additional temporary storage. In this way, hotdeskers managed their information in a wider physical context than simply leaving materials in a traditional office or cubicle workspace. This physical spreading of information within and beyond the confines of the organizational office meant that hotdeskers spent considerable energy managing and coping with their information environments. The need for storage led to greater assertion of ownership over workspaces, with custody of a workspace expressed through such acts as displaying a nameplate or reorganizing shared materials on the hotdesk. Greater nesting occurred when participants stored boxes under the desk in their workspaces, posted notes on the cubicle divider, and left materials on the desk for extended periods of time.

Including Alternative Forms of Working in Our Modeling of Information Behaviour:

Although hotdesking is considered a mode of working that should facilitate mobile work, the information seeking and coping behaviour adopted by hotdeskers in this study showed marked similarities to that discovered previously among home-based eWorkers (Fulton, 2001). Fulton's (2001) study found that eWorkers missed sources, including people sources, when working from home and that they experienced different information gaps from their at-office counterparts. The current hotdesking study shows further how those gaps reveal different barriers to information seeking which are not necessarily resolved by temporary desking.

Figure 1 illustrates the route to information that hotdeskers followed in this study. Hotdeskers experienced imperfect access to information, caused by a lack of or complicated electronic access to information, the need for enhanced technology for seamless connectivity, and lack of storage space. Hotdeskers worked around these barriers to access by gathering and keeping information around them, traveling to information sources, and relying on colleagues for assistance. As instances of hotdeskers circumventing the usual workplace route to information increased, hotdeskers were also more likely

to make that alternate route to task-related information a permanent part of their information seeking. For example, building and maintaining a bank of information in anticipation of future information needs was a very typical means of pre-empting interruptions of information flow by barriers to information, and this collection of information very quickly became a trusted, first choice among resources for information. In this case, hotdeskers actually nearly approach the information seeking process in reverse order, accumulating information they may or may not use in the course of completing a given work task.

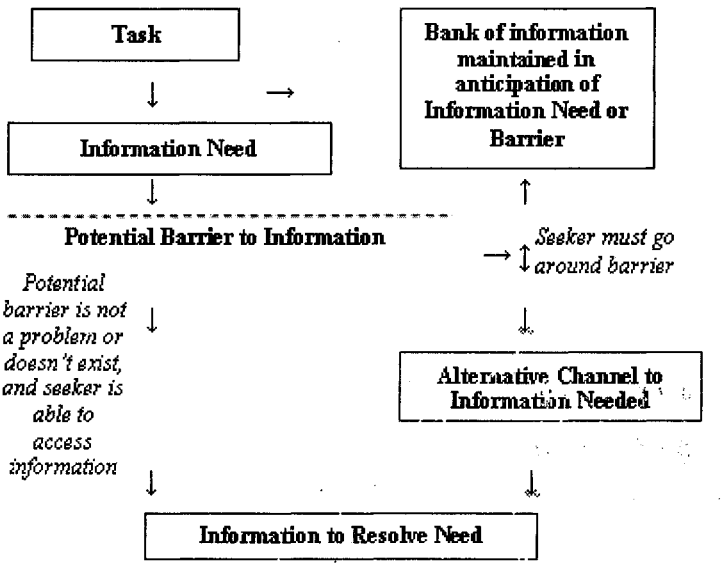


Figure 1. Hotdesker's Information Seeking Behaviour

While it might be argued that general information seeking models provide a looping back to renegotiate or revisit a previous step in the process, it must be noted that hotdeskers are not merely circling back to revise their approach. Hotdeskers may revisit a previous step, but they most often explore a way or ways around the barrier, which stands between them and the information they need or want. An alternate route might include a different channel to or another source of information.

Best Practices for Hotdesking:

Understanding the alternative routes to information taken by hotdeskers is useful in creating strategies for information provision and best practices in alternative work arrangements. In this study, hotdesking was often combined with other forms of eWork, such as working from home or in other remote work locations. For employers, hotdesking offered a means of minimizing real estate, while maximizing use of available space and distributing labour to areas where and when it was needed. For employees, hotdesking provided a workspace in temporary work contexts in the organizational office, as well as in the home and on client sites.

Was hotdesking a link in participants' information chains? Participants reported various ways in which they adapted their information seeking strategies as a result of hotdesking. They reorganized and reformatted information to suit their mobile worklives. They carried information with them to help them complete work tasks and they visited particular work locations where they had stored information to collect that information. Instances of missing information were frequently related to the extent of hotdesking in an organization; indeed, participants who had greater and immediate access to electronic information were less likely to identify serious problems with information seeking and information storage. Participants who relied on a mixture of sources experienced a dispersal of information among storage locations, which, in turn, increased feelings of blurred boundaries between work locations for some participants. On the other hand, hotdesking also provided a vital link between other forms of working, such as working while traveling and working from home, providing a focal point for work, if only temporary.

The hotdesking experiences of both Canadian and Irish participants suggest several best practices in continued hotdesking arrangements:

- Provision of information in electronic format;
- Provision of tools which facilitate seamless connectivity;
- Evaluation of work roles and information access and exchange within those roles to accommodate working in temporary spaces;
- Provision of some form of "home base" for hotdeskers to foster feelings of inclusion and to provide adequate archival space for electronic and print information sources;
- Full, immediate access to work-related information through electronic networks, as well as time-sensitive access to paper-based work information.

Many of these best practices were seen in hotdesking arrangements observed in this study. However, the implementation of these recommendations for hotdesking depended on the extent of hotdesking in particular organizations, and hotdesking arrangements varied widely between organizations. Hotdeskers would benefit from a more systematic

implementation of this work form, incorporating these points for best practice as a foundation for improved future hotdesking.

References

Canadian Telework Association. Available: http://www.ivc.ca/.

Cheuk, B.W. (1998). Modelling the information seeking and use process in the workplace. *Information Research*, 4 (2), 7 pp. Available: http://www.shef.ac.uk/uni/academic/I-M/is/publications/infres/isic/cheuk.htm.

Flanagan, J.C. (1954). The Critical Incident Technique. *Psychological Bulletin*, 51 (4), 327-358.

Fulton, C. (2001). The Case of the Missing Information Resources: Information Seeking in Teleworking Arrangements. *The New Review of Information Behaviour Research* 1, 117-134.

Leckie, G., K. Pettigrew, and C. Sylvain. (1996). Modeling the information seeking of professionals. *Library Quarterly*, 66 (2): 161-193.

Miles, M.B. and A.M. Huberman. (1994). *Qualitative Data Analysis*. 2nd ed. Thousand Oaks, CA: Sage Publications.

Telework in Europe: 2000 Report. http://www.ivc.ca/European.html.

Telework Ireland, 9th National Conference, November 21, 2001, Galway Bay, Oranmore, Co. Galway.

Wilson, T. (1999). Models in information behaviour research. *Journal of Documentation*, 55 (3), 249-270.

DEVELOPMENT OF LOCATION-AWARE APPLICATIONS

The Nidaros framework

Alf Inge Wang[1], Carl-Fredrik Sørensen[1], Steinar Brede[2], Hege Servold[3], and Sigurd Gimre[4]

[1]Dept. of Computer and Information Science, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway, [2]Telenor R&D, NO-7004 Trondheim, Norway, [3]Bekk Consulting AS, NO-0102 Oslo, Norway, [4]CIBER Norge AS, NO-0103 Oslo, Norway

Abstract: This paper presents the Nidaros framework for developing location-aware applications that provide location dependent functionality based on the current location of the user. The framework can be used to develop location-dependent advertisement, city guides, guides for tourist attractions, etc. The framework consists of three main components: *A runtime system* that manages user locations and the interaction with the user clients; *a creator tool* that is used to map information and multimedia content to locations; and *a logging tool* that is used to log the movement of users to monitor the interest for certain locations. The paper also describes an implementation of a location-aware tour guide for the Nidaros Cathedral in Trondheim that can run on different mobile devices. Further, the paper describes experiences from installing, configuring, and running a location-aware tour guide in a real environment. A demonstration of the tour guide was tested on PDAs and mobile phones.

Keywords: Context/location, case studies and experience, applications, tour guide

1. INTRODUCTION

In 2005, it is estimated that there will be more than 1.5 billion wireless subscribers worldwide. Although mobile computing gives challenges to the application developer like handling wireless networks, heterogeneity, limited screen size, input device CPU, memory and battery (Satyanarayanan, 1996), it gives possibilities to develop new types of applications. One such type is location-aware applications. Location-aware applications are useful for

several reasons: Firstly, the limited screen size of mobile devices can be better utilized by providing user interfaces that are related to the context of the user. By using the location, the parts that are not relevant to the current location can be left out. Secondly, the user experience can be improved by providing the user with information and interaction that are relevant to the current context. Here the context is necessarily not only location, but can also be time, weather, temperature, altitude etc. Thirdly, a system can collect context information from users to further improve the location-aware system. For instance in a tour guide system for an art gallery, the system can log which paintings the visitors spend most time by. This information can be used to publish additional information about the most popular paintings in the location-aware guide system.

Several location-aware systems for tour guiding have been developed like the Lancaster GUIDE (Cheverst et al., 2000), CyberGuide (Abowd et al, 1997) and MUSE (Garzotto et al., 2003), but most of these systems are tailored for a specific location-aware application. In 2003, the Norwegian University of Science and Technology (NTNU) started together with Telenor, the largest telecom company in Norway, to develop a general framework for creating, running, and analysing mobile location-aware systems. The motivation for this work was to enable rapid development of location-aware systems that can provide the user with information or multimedia content dependent on the user location. Another important aspect of the framework was to enable support for different mobile clients with different characteristics from the same system. From similar projects, we have found that location-aware systems use various client devices from rather big portable PCs down to small PDAs (Sørensen et al., 2003). Also we noticed that some location-aware systems use customized hardware to get the required characteristics. In addition, the evolution of mobile devices makes it necessary to be able to adapt to future devices with new and useful capabilities. From talking with people managing a PDA-based tour guide (not location-aware) at the Nidaros Cathedral, we understood that theft was a serious challenge. Letting people use their own mobile equipment (like mobile phones) for such services was found to be very interesting.

Another shortcoming for many of the existing systems is that they are tailored to support only one type of positioning technology like GPS, GSM, Bluetooth, IR or WLAN positioning. We used in the Nidaros framework a location server to fetch the user positions from various sources and to send this information back to the system when needed. The location server examines position technologies available to return the most accurate position.

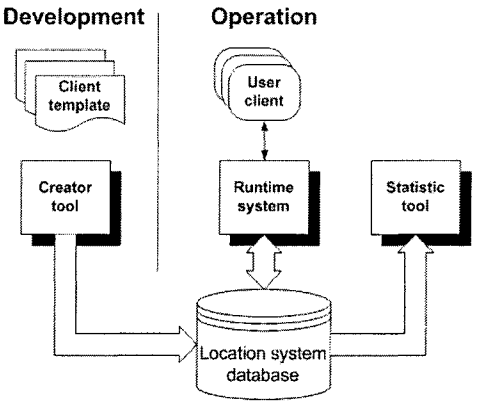


Figure 1. The Nidaros framework for development of location-aware applications

Figure 1 shows an overview of the Nidaros framework for development of location-aware applications. The framework covers development and operation of a location-aware system. The development phase is supported in the framework with a creator tool to add and map location-dependent content, an XML client-server interface and templates for creating clients. A runtime system and a statistic tool support the operation phase. In addition the framework make use of a location system database and a location server.

2. THE FRAMEWORK

This section presents the main components in the Nidaros framework.

2.1 The Development Phase

The development phase of the framework is supported by three components: A creator tool, client templates, and the XML interface between client and server.

The Creator Tool

The creator tool is a system for managing information and multimedia content related to locations that are to be part of a location-aware application. The tool provides a simple user interface that facilitates describing areas hierarchically into maps, zones and objects as a tree structure. A map represents the whole area of the location-aware application. A map can be divided into one or more zones that represent some specific

areas in the map. Within a zone, you can add several objects. These data objects typically represent physical objects and may contain information, audio-clips, video-clips and similar. The maps, zones and objects are mapped to a local Cartesian coordinate where the x- and the y-axis are represented with a 32-bit integer. The coordinate system can be mapped to various geographical positioning systems and makes the framework independent of the actual coordinate systems. By using a Cartesian coordinates in the application, it is easy to visualize maps and objects on the screen of the client device. In the local coordinate system, a map is represented as a rectangle. A zone is represented as a polygon of four coordinates. An object is represented by a specific coordinate and with a hotspot area represented by a polygon similar to a zone. The hotspot area is used to determine if a user is close to an object to trigger some location-aware event.

In the tour guide application we developed for the Nidaros Cathedral (described in Section 3), the map represented the whole cathedral, the zones represented the main parts of the cathedral, and the objects represented physical objects like alters, the pipe organ, paintings, etc.

The creator tool offers a user interface for how various devices like laptops, PDAs and mobile phones are mapped to the local coordinate system. The mapping identifies the type of device by name, the size of the device, an offset coordinate (X_{offset},Y_{offset}), a rotation coordinate ($X_{rotation}$,$Y_{rotation}$) and a rotation angle R_{angle}. These data are used to compute the transformation from real world coordinates to the local coordinate system.

The creator tool can be used to graphically visualize the results of using the tool. The visualization shows the map you have created with named zones and objects. The zones and the objects are shown in different colours. Hotspot areas are shown for the objects.

The Client Templates

The framework provides a template for reducing time to implement clients for location-aware applications. The template is intended for clients that run Macromedia Flash applications. Flash makes it possible to make advanced and dynamic interfaces, and can run on various operating systems like MS Windows, Mac OS, Linux, Unix and Pocket PC. The Flash player is also capable of audio and video playback.

The template offers the basic functionality needed to implement location-aware clients. The template includes functionality for server communication using XML, graphical highlighting of zones and objects, management of hotspots (event-triggered actions), and initialization of multimedia playback. When the template is used, the developer has only to design the user

interfaces and possibly additional functionality if wanted. The template makes the implementation of clients easier and possible misunderstandings of the XML-interface with the server are avoided.

The XML Interfaces

The Nidaros framework provides an open XML interface to the runtime system that makes it possible to create clients for different devices using different technologies like Flash, Java 2 Micro Edition, HTML, .NET compact framework, etc. The only requirement is that the client is capable of managing XML communication and adheres to the specified XML interface. The client application can send a request to the runtime system in several different ways, but it has to follow a predefined XML syntax. The response will likewise follow a predefined syntax. If the client application requires downloading of multimedia content, this is done by an HTTP-request to a file server (see Section 2.2).

The root element in every XML request is a *locationRequest*. This root element can contain several different elements depending on what kind of information that is wanted. Every *locationRequest* must contain an element called *mac* that holds the mac address of the client device. The runtime server needs the mac address to identify and find the position of the user. The following information can be requested from the server: **Position** will get the current position of the user, **simulatedPosition** will get a simulated position of the user (useful for demonstration and testing), **friendsPosition** will return positions of other users registered as friends, **dynamicInfo** will return objects that the user is within the hotspot area of, **tracks** will return some predefined routes that the user might want to follow, and **messages** is used for sending and retrieving messages between users. When a request is sent to the server, the client will get a response from the server with the necessary information depending on the request type.

2.2 The Operation Phase and Runtime System

The main components used in the operation phase are the user clients, the runtime system and the location system database. The user clients are developed using the client template and the XML-interface described in Section 2.1. The runtime system is the heart of the Nidaros framework that brings location-aware applications alive. The runtime system manages the information in the database including maps, zones, objects, users, and etc. The main task of the runtime server is to feed the clients with correct information according to the client position.

Figure 2 shows the physical view of the runtime system. The runtime system supports several types of clients and identifies three client types we have implemented. The figure shows that wireless LAN is used between the server and the clients, but also other types of wireless networks have been used. Currently, our mobile phone client uses GPRS for communication between the client and the server. The runtime system itself consists of four main components described below.

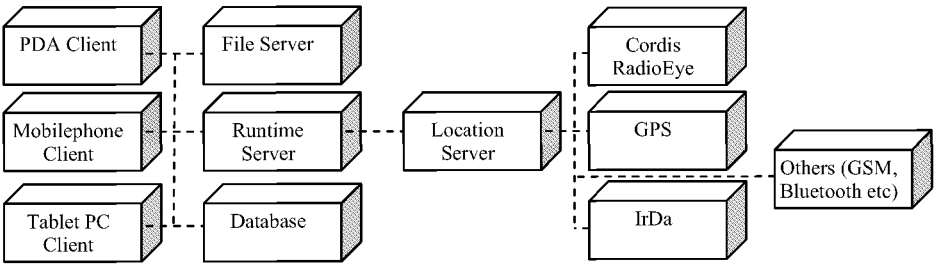


Figure 2. The Physical view of the runtime system

The **file server** stores media files accessible for mobile clients. A file server is used because mobile devices are not likely to store all media files locally because of the limited memory. How much the file server is used depends on the magnitude of multimedia used and the storage capability of the mobile device. Multimedia elements used often should be cached on the device.

The **runtime server** is responsible for handling requests from the clients and providing the services requested. A more detailed description of this component is given later in this section.

The **location system database** stores all information used by the clients and the server. This database is also used by the creator tool when creating location-aware applications, and by the statistical tool for analysis of user patterns.

The **location server** gets the current positions of the clients through various interfaces to different position technologies like WLAN positioning, GPS, IrDA, Bluetooth etc. A more detailed description can be found in Section 2.4.

From early scalability tests we have found that the wireless network between the clients and the server will be the main bottleneck of the system. If the GSM network is used, only audio streaming is supported. If WLANs like IEEE 802,11b are used, video streaming is supported. The number of simultaneous users to be served depends on how well the physical network is implemented. Another possible bottleneck can be the file server. However, such servers can be duplicated to achieve better performance.

Figure 3 shows the logical view of the runtime server architecture. The system is communicating with client applications through the *GLocServlet* class. Data is exchanged as XML, and the *XMLTransformer* class interprets and transforms the information sent between clients and the server. For the runtime server, we decided to use an architecture based on a centralized control model. This means that all information flow through the *MainController* class. By using centralized control, it is easy to analyse control flows and get the correct responses to the given client requests. It also makes it easy to substitute the servlet class with another class for handling the client communication and to add new interfaces to the system as needed. The *UserManager* class is responsible for maintaining information about the users. This task includes storing the user's last position and deciding whether a person is allowed to communicate with another person (defined as friend). The *PositionManager* class is responsible for returning the user position, adjusted to the type of mobile device used. The *TrackManager* class is responsible for keeping information about the available predefined tracks. Each track has a unique name, so the client application can either request all tracks or one particular track. The *DbManager* class is responsible for all communication with the database.

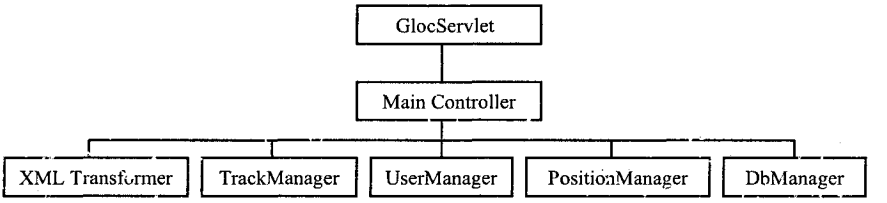


Figure 3. The Logical view of the Runtime system

2.3 The Analysis Phase

The statistical tool is useful for analysing the use of location-aware applications. The tool uses data logged by the runtime system to look at user behaviour and what objects that are most popular. Four classes implement the statistical tool: GUI, LogTool, DbManager and DbLog. The GUI class presents the different services. The LogTool class is responsible for calculation and manipulation of the data stored in the database. The DbManager and DbLog classes handle database issues.

2.4 The Location Server

The location server (see Figure 4), developed by Telenor R&D, is a framework for uniform, geometric-based management of a wide variety of

location sensor technologies. The goal of this framework is to have one server that can get positions of mobile devices through multiple location sensing technologies. By using the location server in the Nidaros framework, we do not need to tailor our system to use a specific positioning technology. We can also use different positioning technologies within the same application.

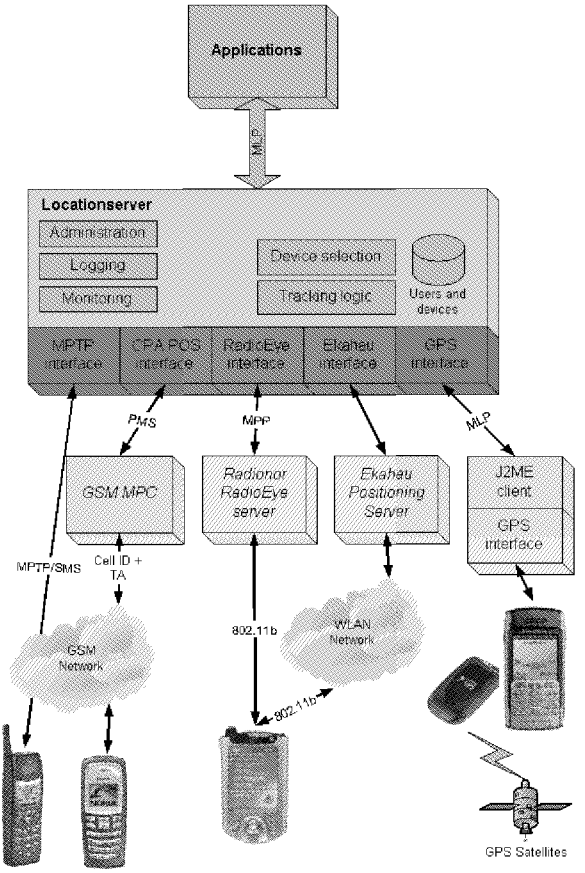


Figure 4. The architecture of the location server

The server includes a middleware protocol specification and a specification of quality-of-service parameters. Further, the server has support for event-driven position reporting (i.e. for change of position) and support for methods for merging and position enhancements. Figure 4 shows an overview of the location server architecture. The architecture is layered and shields the application from the details of collecting and merging location information from a variety of sources. The devices that are positioned by the server are identified by the MAC-address of the device. The location server also manages authorisation for accessing the position of a device. Further,

the architecture provides support for monitoring and tracking of mobile devices. The architecture provides several interfaces to various positioning technologies. The location server currently supports positioning using GPS, GSM, and WLAN.

One advantage from using the location server is that this server handles the complexity of merging locations that might include partly overlaps of positioning systems and seamless transitions between different position systems. This is especially useful for location-aware applications that cover both indoor and outdoor areas.

2.5 The Location System Database

The location system database stores location data, log data, and user data. The *location* data is represented in five tables describing maps, zones, objects, preferences, and mapping. The motivation for these tables is that one location can have several maps covering different territories, each map can cover several zones, each zone can contain several objects, and each object can be connected to several preferences. The preferences of an object are used to provide the dynamic menu service. The user can state the preferences in the attraction, and only objects matching his preferences will be displayed in the client menu. In addition, there must exist a table with mapping information to transform locations from world coordinates to coordinates adjusted to fit the client map.

The *user* data is represented in three tables describing users, user groups and user preferences. The user table contains an MAC-id of the user device and other information. The *user group* table is used to group users that are friends to allow services like tracking friends and messaging. The *user preferences* table stores information about the user's main interests.

The *log* data is represented in two tables. One table is used for storing user movements and one table for storing what kind of objects the user is interested in. The statistical tool uses the log data.

3. IMPLEMENTING A TOUR GUIDE USING THE FRAMEWORK

We created a location-aware tour guide for the Nidaros Cathedral in Trondheim to test the Nidaros framework in a real setting. The cathedral had an existing PDA-based tour guide, called Nidaros Pocket Guide (NPG) that was not location-aware. To show the flexibility of the Nidaros framework,

we decided to implement support for two different types of clients: A PDA and a mobile phone.

3.1 The PDA Client

We decided to develop our location-aware PDA client in Flash MX. This made it possible to reuse code from the NPG and the client template. Our PDA application provided functionality for selecting language, enable trace of own movement, show friends on a map, show zones and objects on a map (with highlighting of current zone and objects), displaying text about objects, and playback of audio or video about an object. The PDA used to run the client was an iPaq with Windows CE and the wireless LAN IEEE 802,11b network integrated. This made it possible to get the position of the device using WLAN-positioning. The location-awareness was presented in the client application in two ways. The first way was to show a map with the position of the user, position of possible friends, highlighting of current zone and nearby objects. The second way was optional for the user, and made the application able to initiate display of objects (text, audio or video) when the user entered the hotspot area of an object. A PDA client is shown to the left in Figure 5.

Figure 5. The system running on iPaq and SonyEricsson P900 and the WLAN tag

3.2 The Mobile Phone Client

A mobile phone client could either be implemented in J2ME or using HTML and the web-browser installed in the phone. We decided to do the latter by implementing an additional servlet component that communicates with the runtime server and produces HTML for the mobile phone. We used WAP push to send the appropriate web pages when the user was within a specific area to enable the client to react on the user location. This made it

possible to, e.g., get the phone to initiate playback of a video when the user was close to a specific alter. We used a SonyEricsson P900 in the test because of its support for WAP push and the built-in multimedia player. To the centre in Figure 5, the mobile phone client is running on a P900. The most common way to position mobile phones is to use GSM positioning. This method is too course-grained to be used for positioning inside a cathedral. To solve this problem, we came up with the idea of letting the user wear WLAN tags that could be positioned using WLAN positioning. WLAN tags are produced by RadioNor Communications, and are small WLAN radios that transmit a MAC-ID. A picture of a WLAN-tag is shown to the right in Figure 5. The size of the tag is about 4cm wide and 3cm high. For the mobile phone client, it was necessary to register the MAC-ID of the WLAN tag and the mobile phone number to link the phone to the tag.

3.3 The Position Technology Used

We used the Cordis RadioEye WLAN positioning produced by RadioNor Communications to position the client devices used in our location-aware application. The RadioEye is a small box with advanced antenna technology and a Linux-server that can determine the physical coordinates of every WLAN-terminal that are active within its coverage area. The sensors decode the MAC addresses of the network devices and determine their geographical position with a typical accuracy of 1-2 meters. A RadioEye covers a radios about 60 degrees from the centre and can e.g. be placed in the ceiling of a building.

3.4 Setting up and Running the Location-aware Application

The demonstration of the location-aware tour guide for the Nidaros Cathedral was performed May 14th 2004. Before we could start to run the demonstration, we had to install the infrastructure needed for running the system. Four RadioEyes were installed at the gallery of the cathedral to cover four different zones of the building. After the RadioEyes were in place, the coordinates from the RadioEye server had to be aligned with the coordinates used by the system. This was done by taking measurements from different locations in the cathedral. These measurements were made in a pre-test before the real demonstration. Approximately 20 people were present at the demonstration representing various technology-oriented companies as well as the media (television, magazines and newspapers). The demonstration of the system was very well received and especially the

mobile phone application attracted much attention. As far as we know, there are no similar location-aware applications running on mobile phones.

4. EXPERIENCES

This section describes experiences we gained from developing a location-aware application using the Nidaros framework. We have found the framework very useful for several reasons. Firstly, the server side of the system can be used directly without any modifications. The only thing missing is the information to be used in the location-aware application. This information can easily be added by using the creator tool. In addition, the creator tool can be used to make changes to the location information before and during the operation phase.

By using the available client template, the time to implement a client is rather short. Currently, client templates are only available for HTML and Flash. It would have been useful to provide templates for Java 2 Micro Edition and .Net Compact Framework. It does not require much work to write a client from scratch using the defined XML-interface. XML parsing might cause a challenge for a client implemented in J2ME because of the memory limitation in J2ME. However, the XML messages used in the Nidaros framework are relatively small and simple, and do not contain several levels. A possible extension of the Nidaros framework could be to make a client generator to ease the implementation of new clients for all client technologies.

The simulated movement of users was found to be a very useful feature of the runtime system and was invaluable for testing client applications. It takes several hours to set up a real environment with real sensors, and it would be time consuming to engage real users to debug the application.

The database used in the framework was found general enough for the tour guide application. However, there are limitations on what type of information that can be stored. This means that the database scheme might have to be extended to fit any location-aware application.

We introduced a file server that could feed the clients with multimedia contents because the limited storage available on client devices. In an ideal system, all the media files should be stored locally on the device for quick and responsive presentations. This was impossible for the tour guide for the Nidaros Cathedral if more than one language should be supported on the same device. Most of the multimedia files were audio files, but it was not storage space enough for more than one language. By introducing more videos, this would be a bigger problem. We found from the demonstration in the cathedral that the user had to wait from 5 to 7 seconds before the audio

or video was played. We stored the most used media files on the device to avoid such long waiting times. An extension of the Nidaros framework could be to have smarter media file communication management. This means, e.g., that the mobile client can start to pre-fetch files that are likely to be played in the near future based on the location and movement of the user.

The mobile phone client got tremendous attention when we demonstrated the location-aware tour guide in the Nidaros Cathedral. The main reason was to be demonstrated such advanced applications running on devices that many own. The current solution involves a WLAN-tag to make the system work. For a commercial tour guide for mobile phones, the WLAN-tag could be available in the entrance of a sight by paying the entrance fee. Sending an SMS that includes the ID of the tag to the tour guide server could initialize the setup of the mobile location-aware tour guide. The combination of using a WLAN-tag together with a mobile phone gives other new exiting opportunities for location-aware applications that can be used both indoors and outdoors. The WLAN-positioning technologies like the RadioEye can be installed in all public building like airports, hospitals, shopping centres etc.

The use of a location server makes it easy to adapt to new positioning technologies when they are available. The only required change of the system is to implement a new interface for the new position technology in the location server.

The choice of using XML for data exchange between client and server has many benefits. The main benefit is the extensibility of the interface and provision of an open interface to other systems. The main disadvantage of using XML is the overhead sending messages between client and server. For a system with many users, this can cause scalability problems because of the limited bandwidth in wireless networks. A problem we experienced running the location-aware tour guide, was the high demand on CPU and memory on the mobile clients for parsing the XML-data. Generally, the clients spent more time on parsing XML data than they used to send requests and receive responses. A high demand on the CPU will also make the battery run out faster. We foresee that this problem will not be so dominant for future mobile devices with improved performance and battery technology.

5. RELATED WORK

This section describes work on similar frameworks from location-aware and context-aware applications.

The NEXUS Project (Volz and Sester, 2000) has developed a generic infrastructure that can be used to implement all kinds of context-aware applications both for indoor and outdoor services. The NEXUS clients

access the server via a standardized user interface running on the mobile device carried by the user. The interface has to be adjusted to the different kinds of clients, especially concerning the different level of computing power, different amounts of memory and different size of displays. A NEXUS station can manage sensor systems that measure both global indicators (like temperature) and object related information (like location). The NEXUS framework uses separate components for sensor management and communication, and use spatial models with multiple representations to model the physical world stored in distributed databases.

The Framework for Location-Aware ModElling (FLAME) is a configurable, generic software framework for development of location-aware applications (Coulouris et al., 2004). FLAME provides support for multiple sensor technologies, provides a simple spatial model for the representation of locatable entities. In addition, FLAME provides a simple event architecture for the presentation of location information to applications, and a queryable location database. The framework and its applications are largely event-driven in order to accommodate the real-time nature of the location information that they handle. The database holds the initial states (like static regions), and it also holds a synopsis of the real-time location information. A region manager stores and retrieves regions from the database. A spatial relation manager generates application-related events to satisfy currently active subscriptions. The event adapters generate events, e.g., when a person has moved a ``Person Movement Event'' is generated.

Dey and Abowd (Dey and Abowd, 2001) presents requirements and a conceptual framework for handling context information. The requirements to be fulfilled by the framework are *Separation of concerns*, *Context interpretation*, *Transparent, distributed communications*, *Constant availability of context acquisition*, *Context storage and history*, and *Resource discovery*. The conceptual framework for handling context by Dey and Abowd suggests the use of **context widgets** to provide access for applications to context information from their operating environment. The context widget is regarded as a mediator between applications and the operating environment, insulating applications from context acquisition concerns. Context-specific operations are addressed in the framework by four categories of components: interpreters, aggregators, services and discoverers. The framework defined by Dey and Abowd focuses more on the management of various context sources and to represent these context sources in an application.

6. CONCLUSION

Although there exist different types of location-aware applications, there are still many location-aware services that have not been explored. Many of the existing location-aware applications are tailored just for one purpose. In this paper, we have presented the Nidaros framework to be used to implement location-aware applications. Our framework provides support for the development phase, the operation phase and the analysis phase. In the development phase the creator tool is used to add needed information into the database to be used by the final application. Further, the client templates can used for faster development of mobile clients with some basic functionality. In the operation phase, the runtime system manages all interaction between clients and the server including communication of multimedia files and determination of clients' positions using a location server. For the analysis phase a statistic tool can be used to analyse the usage of the location-aware application, detect possible bottlenecks of the system, and see what objects that are most popular.

References

Abowd, G. D., Atkeson, C. G., Hong, J., Long, S., Kooper, R., and Pinkerton, M. (1997). Cyberguide: A Mobile Context-Aware Tour Guide. Wireless Networks, 3(5):421–433.

Cheverst, K., Davies, N., Mitchell, K., and Friday, A. (2000). Experiences of developing and deploying a context-aware tourist guide: the GUIDE project. In Sixth Annual International Conference on Mobile Computing and Networking, pages 20–31, Boston, Massachusetts, United States. ACM Press.

Coulouris, G., Naguib, H., and Samugalingam, K. (2004). Flame: An open framework for location-aware systems. http://wwwlce.eng.cam.ac.uk/qosdream/ Publications/flame.pdf. Submitted for publication.

Dey, A. K. and Abowd, G. D. (2001). A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications. Human-Computer Interaction (HCI) Journal. Special Issue: Context-Aware Computing, 16(2–4):97–166.

Garzotto, F., Cinotti, T. S., and Pigozzi, M. (2003). Designing multi-channel web frameworks for cultural tourism applications: the MUSE case study. In *Museums and the Web 2003*, Charlotte, North Carolina, USA.

Satyanarayanan, M. (1996). Fundamental Challenges in Mobile Computing. In Fifteenth Annual ACM Symposium on Principles of Distributed Computing, pages 1–7, Philadelphia, Pennsylvania, United States. ACM Press.

Sørensen, C.-F., Wang, A. I., and Hoftun, Ø. (2003). Experience Paper: Migration of a Web-based System to a Mobile Work Environment. In IASTED International Conference on Applied Informatics (AI'2003), pages 1033–1038, Innsbruck, Austria.

Volz, S. and Sester, M. (2000). Positioning and Data Management Concepts for Location Aware Applications. In 2nd International Symposium on Telegeoprocessing, pages 171–184, Nice-Sophia-Antipolis, France.

DECOUPLING DESIGN CONCERNS IN LOCATION-AWARE SERVICES

Andrés Fortier, Gustavo Rossi, and Silvia Gordillo
LIFIA. Facultad de Informática. UNLP. La Plata, Argentina
{andres,gustavo,gordillo}@lifia.info.unlp.edu.ar

Abstract: In this paper we present an original approach to design and implement applications that provide location-aware services. Our approach emphasizes a clear separation of the relevant concerns in the application (base behavior, context-sensitive properties, services, etc.) to improve modularity and thus simplify evolution. We first motivate the problem with a simple scenario of a virtual campus; we next discuss which are the most important concerns in the application, we explain why we must separate them and show a simple approach to achieve this separation. We analyze the most important (sub) models in which we decompose a location-aware application and explain the use of dependency mechanisms to trigger behaviors related with the provision of services according to the user position. We briefly describe a proof of concept by means of an archetypical implementation we developed following our ideas. We next compare our work with others and discuss some further work we are pursuing.

Keywords: Location-aware services, location sensing, concern decoupling, modularity

1. INTRODUCTION

Context-Aware (and in particular Location-Aware) applications are hard to build and more difficult to maintain due to their "organic" nature (Abowd, 1999). For this reason, improving modularity is extremely necessary when designing this kind of software. Dealing with location (and other kind of context) information is essentially hard because this information has to be acquired from non-traditional devices and distributed sources, and it must be abstracted and interpreted to be used by applications (Dey, 2001).

While much research on context-awareness has focused on solving these problems, and many Context-Aware (CA) applications and frameworks have been built in the last years (Bardram, 2005; Hofer et al., 2003; Salber et al.,

This paper has been partially supported by the Argentine Secretary of Science and Technology (SeCyT) under the project PICT 13623

1999), there is still a poor characterization of those software design issues that make CA software difficult to build. In addition, CA applications have to deal with the following problems:

- Abstracting context means more than changing representation. Even though it is clearly explained in (Dey, 2001), the process of context interpretation usually ends far from application concerns. While interpreted context data is usually dealt as strings, applications are composed of objects, which means we have to deal with this impedance mismatch.
- Adapting to context is hard; design issues related with context-aware adaptation are not completely understood and thus handled incorrectly. For example, although rules can be useful (especially if we want to give the user the control of building his own commands), we claim that more elaborated structures are needed to improve maintenance and evolution.
- Context-related information is usually "tangled" with other application behavior. For example, the location of an application object (which is necessary to detect when the user is near the object) is coupled with others object's concerns, making evolution of both types of characteristics difficult.

Our research deals with the identification of recurrent problems and design micro-architectures in CA software. In (Rossi et al., 2005) we argued that design patterns are an excellent way to record and convey design experience related with CA (Abowd, 1999) adaptation. In this paper, we go further and describe an architectural approach for dealing with the problem of providing CA services (Bardram, 2005). Our approach is based on a clear separation of concerns that allows us not only to decouple context sensing and acquisition (as in (Salber et al., 1999)), but mainly to improve separation of application modules, to ease extension and maintenance. For this purpose we make an extensive use of dependency (i.e. subscribe/notify) mechanisms to provide context-aware services.

Along the paper we will show how to separate application concerns related with context awareness to improve modularity and, as a by-product, we will present a strategy to extend legacy applications to provide location and other context-aware services. In order to be consistent, we will treat services as full fledged objects and make them dependent of context changes.

The rest of the paper is organized as follows: In Section 2 we introduce a simple motivating example both to present the problems and to use it throughout the paper; in Section 3 we describe the most important concerns in this kind of software and introduce our criteria to decompose the application into layers and components. A complete description of each of the different models comprising our architecture is shown in Section 4. In Section 5 we briefly

describe an archetypical implementation. In Section 6 we compare our work with related work in this field and finally, in Section 7, we conclude and discuss some further work.

2. MOTIVATING EXAMPLE

Suppose we are adapting an existing software system in a University Campus to provide context-based services (in particular, location-based ones), in the style of the example in (Sousa and Garlan, 2002). Our system already provides information about careers, courses, professors, courses material, timetables, etc. We now want that users carrying their preferred devices can get information or interact with the system while they move around the campus. For example, when a student enters a classroom, he can get the corresponding course's material, information about its professor, etc. At the same time, those services corresponding to the containing location context (the Campus) should be also available. When he moves to the sport area, the course's related services disappear and he receives information about upcoming sport events and so forth. It should be noticed that different contextual information such as the user's role or activity might also shape the software answer.

The first design problem we must face is how to seamlessly extend our application in order to be location-aware, i.e. to provide services that correspond to the actual location context. The next challenge involves adapting the behavior to the user's role (a professor, student, etc) and other meaningful contextual parameters such as current time or user's activity. While applications of this kind have always been built almost completely from scratch, we consider that this will not be the case if context-aware computing becomes mainstream; we will have to adapt dozens of legacy applications by adding new, context-aware behaviors.

When working with CA applications we can find typical evolution patterns such as adding new services related to a particular location, improving sensing mechanisms (for example moving from GPS to infrared), changing the location model (from symbolic to geometric), and so on. While most technological requirements in this scenario can be easily fulfilled using state-of-the art hardware and communication devices, there are many design problems that need some further study. The aim of this paper is to focus on a small set of those problems, mainly those that characterize the difficulties for software evolution. We stress on those features specific to this particular example because it is a good stereotype of a family of software applications with similar problems.

3. IDENTIFYING AND SEPERATING DESIGN CONCERNS

As previously mentioned, well-known approaches to context-aware applications design have clearly identified some broad concerns that must be separated for achieving modularity: sensing (implemented for example as Widgets in (Dey, 2001)), interpretation (also mentioned as context management in (Hofer et al., 2003)) and application. Layered architectural approaches such as in (Hofer et al., 2003), or MVC-based ones like (Salber et al., 1999) provide the basis for separating those concerns using simple and standard communication rules. However, applications (the third concern) are considered as being monolithic artifacts that deserve little or no attention. It is easy to see in the motivating example that the gap between application objects (in particular their behaviors) and the outer (context-related) components is not trivial. Of course, one could argue that once captured and interpreted, context information is not different to other "old-fashioned" application data, and thus we can use the very same techniques, which allowed us to survive in the past when dealing with input information. As a simple counter example let us take into account location data: to check that a user is in a campus' room, we must compare his position with the room location; is this location an attribute of the room object? What happens if we use different location models? Should we clutter the room object with these variants? Moreover, suppose that we are adding location-aware functionality to an existing system; should we change the base application behavior and write the code for providing location-awareness inside application objects? Following this thread, we may ask ourselves how to cope with services: are they supposed to be application behaviors (i.e. should we consider the services as methods of the room object?) or should they be decoupled into independent objects? The same problems also appears when dealing with other contextual information that cross-cut application objects.

In our research we have identified a set of concerns that should be clearly separated to improve evolution and maintenance: applicative, location and service concerns should be as independent as possible.

In the rest of the paper we will elaborate our strategy for building location-aware software and we will describe the previously mentioned concerns and how they interact with the lower-level ones (such as the sensing concern).

4. DESIGNING LOCATION AWARE SERVICES

In the following sub-sections we assume that we need to extend an existing application with location-aware services. This application implements the base behaviors on top of which services are built. For the sake of understanding we first describe the overall architecture and concentrate later on each software component. The preceding example is used throughout the paper.

4.1 The Overall Architecture

To improve the description of the important architectural decisions, we present two orthogonal views showing different design concerns and how they relate with each other: an application-centered view and a sensing view.

Application view. This view (shown in Figure 1(a)) concentrates on the application model. In the first layer we specify the application model with its "standard" behaviors; application classes and methods are not aware of the user's context. In our example we would have classes to handle room reservations, professors and material associated with each course, etc. Note that in this layer the concept of a "user" does not exist, though we might have objects that correspond to different user roles, such as students, professors and so on.

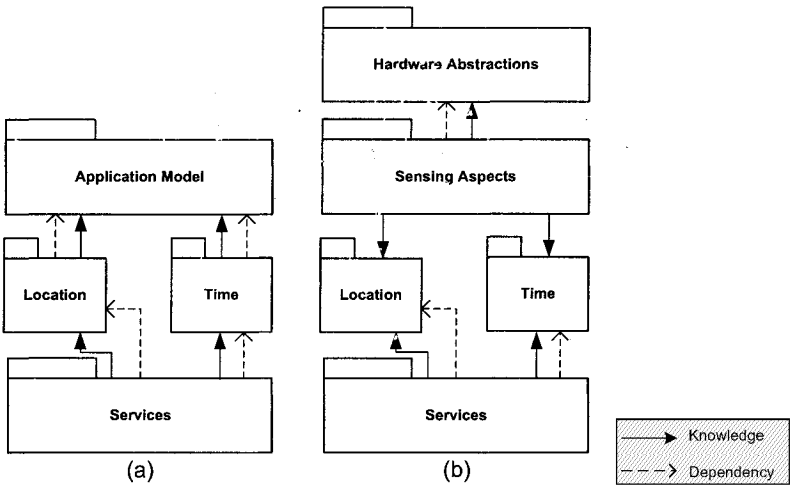


Figure 1. A layered architecture for Location-Aware Services

The second layer contains a set of components that extend the application model with information needed to provide context-aware behavior. For example, the campus, the sport field and the rooms have an associated location that is used to determine if the user is inside one of those areas. It is important to notice that Location objects do not belong to the application concern; according to our approach the basic behavior of a room should not be cluttered with geographic information. As described in section 4.3, decoupling location from other application objects allows us to deal with different location models transparently.

Finally, the third layer contains the (location-aware) services. These services are modeled as objects that will be further associated to certain geographic areas by means of a subscription mechanism.

Relationships among objects in different layers follow two different styles: typical knowledge relationships (such as the relationship between the object containing a room's location and the room itself) and dependency relationships (in the style of the Observer pattern (Gamma et al., 1995)) that allow broadcasting changes of an object to its dependent objects. In Figure 1(a) we also show an additional Package (Time) as an example of other context-related modules that may be included in the second layer.

In Figure 2 we show a small example exploiting the packages in Figure 1(a). Classes like `Course`, `Teacher` and `Room` belong to the application model and have no location-related behaviors. Location-aware classes (in the bottom of the diagram) "observe" application model classes and add additional context behavior. Notice that in this layer we introduce the notion of a `Location.User`, i.e. the location aspect of the user of our context-aware application. This aspect relates with the actual user's role through a `Person` instance.

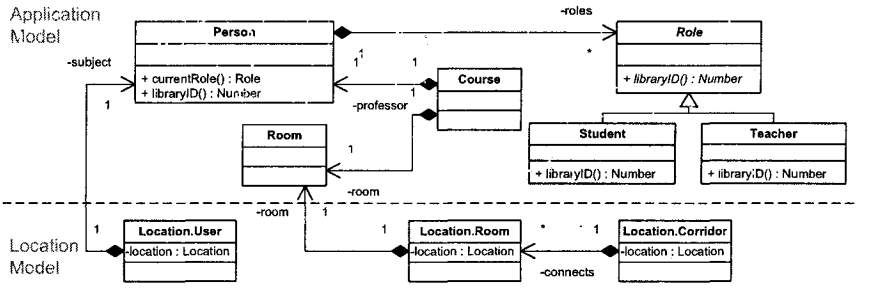


Figure 2. Location-Aware Classes vs. Application Classes

Sensing View. In Figure 1(b) we present another architectural view of our approach. In the first layer we find the hardware abstractions used for gathering data, such as IButton, InfraredPort, GPSSensor, and so on; this abstractions have some points in common with Dey's Widget components (Dey, 2001).

The second layer comprises higher level sensing aspects implemented as objects that plug the lower level sensing mechanisms (in the hardware abstraction layer) with the aspects that are relevant to the application's context that have to be sensed. This decoupling guarantees that the location model and the sensing mechanisms can evolve independently. For example, we can use a symbolic location model (Leonhardt, 1998) to describe locations, and infrared beacons as sensing hardware; we can later change to a non-contact iButton seamlessly by hiding this evolution in the sensing layer.

Modeling and describing the user. As shown in Figure 1(a) we decided to model each context concern in a separate package. This idea is also applied to model the user: we consider the user model as being composed of different aspects, each one acting (differently) on the services that are available to the user. In our example, these services depend on the user's location and thus we need to model a user's aspect that handles the location concern. If, in the future, we decide that the way in which services are presented to a user may also depend on his preferences (explicitly stated by him or inferred from his usage history) we will need to add a new view which handles this aspect. Once the different concerns are modeled, we need some object to coordinate all views and decide how changes affect the user's services. We decided to design this coordinator in the service layer. This object, from now on called user object (or `Service.User`) knows, and it is dependent of, every concern that affects the user and reacts based on those concern's changes. The user model can be thought as cross-cutting the Services and Location layers; it comprises packages that belong to each of them.

4.2 Application Layer

In our architectural framework, the application model contains those classes specific to the intended domain and whose behavior do not depend on contextual information. In Figure 3 we show a simplified class model for the exemplary application. Notice that, for example, a room can return the courses occurring on that room at a particular time; a course meanwhile can provide its content material, the list of enrolled students, and so on. Also, the different roles modeled in the application might be eventually used for role-aware services. Notice also that there is no information on location, space, etc.

4.3 Location Layer

In the location layer we design components that seamlessly "add" location properties to those objects (in the application model) that must "react" when the user is in their vicinity. For example, to be able to say that a user is in room "A" we first need to create a location abstraction of the corresponding room object. By clearly decoupling the location from the application object we can use different location models (Leonhardt, 1998) in an unobtrusive way. In Figure 4 we show the class diagram of a simple location "map" of the campus. Note that the location layer also comprises classes for "pure" location concepts; for example corridors and maps don't have a counterpart in the application layer. In our example, we may be interested in representing a map of the university building, where we find rooms that are connected by corridors.

To achieve higher levels of reuse, we further decouple location objects from the specific location model we use for them. As a result of this separation, we

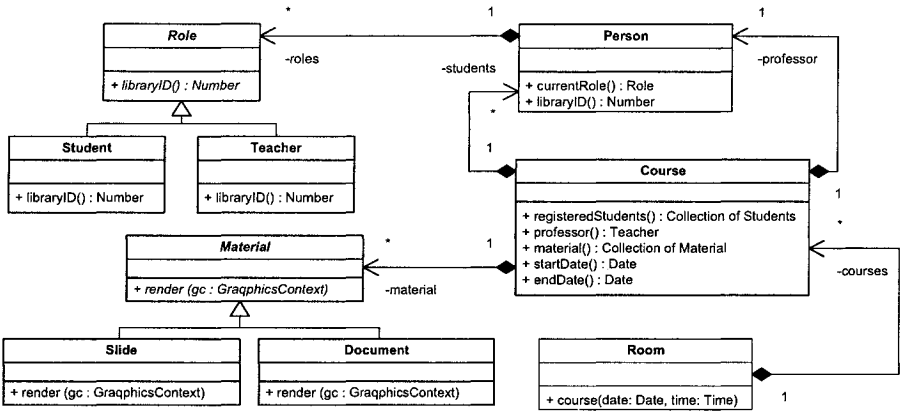


Figure 3. Class Diagram of the University Campus

end up with location objects (like rooms and corridors) that are aware of having a specific location, but that are independent of the location model being used. This independence is achieved trough the `Location` interface, which specifies the basic behavior that every location model should implement. Using this approach, implementation details (for example, knowing if a location is inside another) are hidden in each location model and allows us to change between location models dynamically without any impact on the system.

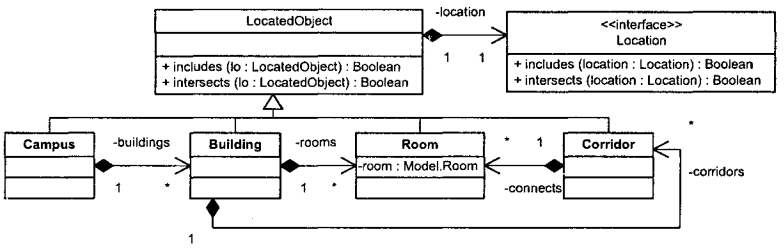


Figure 4. Class diagram of the University Campus Location package

4.4 Service Layer

We consider (context-aware) services as possible independent artifacts which are developed individually and do not need to interact with each other. We also view them as extending some existing application behavior and thus they might need to interact with application objects. Also, service users are immersed in a

service environment, which reifies the real-world environment. A `Service.User` (i.e. the user considered from a services point of view) is modeled so that we can reflect those services to which a person is subscribed to, which services are currently available and so on. The `Service.User` object is also used to build the whole picture of a user, mediating between every possible context aspect that is relevant to that person. In this layer, the `Service.User` knows (and is dependent of) a `Location.User`, so that the service layer can react to changes in the location layer. In the remaining sections of this paper, each time we talk about a user we will be referring to a `Service.User`.

The service environment is in turn responsible for handling available services, configuring service areas and mediating between users and services. A service may be as simple as an alarm (that is triggered when we enter a place at a certain time) or as complex as a full-fledged application. Services are modeled as first-class objects which share a common super-class (or implement a given interface); this allows our framework to treat them uniformly and simplify the addition of new services. In the following sub-sections we give a brief outline of how services are modeled and implemented using this approach.

Creating New Services. New services are defined as subclasses of the abstract class `Service`, playing the role of a Command (Gamma et al., 1995). The specific service's behavior is defined by overriding appropriate methods of `Service` such as *`start()`* (used to perform initialization stuff), *`activate()`* (triggered when the users selects the service form the available services list), etc. In our example, the `CourseMaterial` service is defined as a sub-class, and the message *`activate()`* is redefined so that a graphical interface is opened to display the courses material. Once the service class has been created and its behavior defined, it has to be published to allow users to subscribe to the service; the *`addAvailableService`* message is used to inform the environment about the new service.

Subscribing to Services. Users can access the available services and decide to subscribe (or unsubscribe) to any of them. The details of the subscription mechanism are beyond the scope of this paper; however, is important to mention that a service can be customized by its user.

Once a user is subscribed to a service and provided he satisfies the service's constraints (for example in relationship to the user's role), he can use the service when entering the area associated with the service.

Service Areas. A key aspect in our approach is that services are associated with (registered to) specific areas, called service areas. When the user enters a service area, all services registered to the area (to which the user has subscribed) are made available. Service areas are defined to achieve independence

from the sensing mechanism. To illustrate the idea, suppose that our sensing mechanism is based on infrared beacons. Since a beacon's signal range is limited, we may need to use more than one beacon to detect the presence of a person in a certain area. As an example, suppose that two beacons (B1 and B2) are placed in the opposite corners of a room (Room A) to detect the user presence. Even when there is a clear distinction between capturing B1's id and B2's id from the location-model point of view, this difference should be transparent to services allocated to the Room (area) A.

Services are not associated to physical areas (in terms of location models) but to logical areas named *service areas*. In this way, we can think of the services that are available in Room A or in the hall, instead of thinking about the services that are triggered by a group of beacons. In Figure 5 we show a class diagram indicating the relationships between the Environment, the Service Areas and the associated Services.

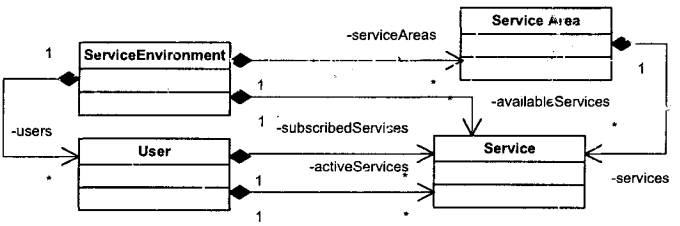


Figure 5. Services and Service Areas

Service Activation. When the person's movement is captured by a sensor, it sends the `location(newLocation)` message to the `Location.User` corresponding to the actual user. This message triggers a change in the location model that is captured (by means of the dependency mechanism) by the `User` object in the service layer. This object interacts with its environment to calculate, based on the user's old location, if the user left a service area. If this is the case, the user object is told to leave that service area by means of the `leaveArea(aServiceArea)` message, which will remove the services provided by that service area from the user's active services. In a similar way, according to the new user's location, the environment checks if the user has entered a new service area. In that case, the user object receives the `enterArea(aServiceArea)` message in order to add the corresponding services.

4.5 Sensing Concerns

We introduce the idea of a sensing concern to separate the context model from the way it is sensed. A sensing concern represents the "glue" between the different aspects that are relevant to our context-aware application and the way they are sensed. A sensing concern is created and configured to be an observer of one (or more than one) sensing mechanism. When a sensor indicates that an event occurred (i.e. some context information changed), the sensing concern acts on its subject by sending an appropriate message.

In this layer, the core behavior is modeled in the `SensingConcern` class and its subclasses. A sensing concern is attached to a sensor with a fetch policy suited for it; for example, a GPS system may need a pull policy while a bar-code reader a push one. Additionally, we specify the message that should be sent to the object that models a specific context concern in order to update its aspect (in our example the *`location(newLocation)`* message should be sent to the `Location.User`). Depending on the programming environment used, this behavior can be achieved by sub-classing `SensingConcern` or via reflection.

Continuing with our example, when the student enters Room A the infrared port of his PDA captures B2's id, and the port abstraction (in the hardware layer) reacts by notifying its dependents that a new id has been received. Since a sensing concern has been created to modify the user's location, it receives the notification and reacts by adding the beacon's covering area to the user's active areas. Once this happens the corresponding user object interacts with its environment to find out which new services are active and available.

4.6 Putting all Things Together

In order to clarify the objects interactions occurring in our architecture, in Figure 6 we present an interaction diagram that shows how a change in the location layer triggers the service assignment to a user. From the services point of view, a change can drive the framework to add or remove service areas depending on the user's previous and current location. To keep the diagram simple, we assume that the initial interaction begins with a message sent by an object of the Sensing Aspects layer. When the sensing hardware (whatever it is) detects the presence of a user in a room, the sensing concern attached to it sends the message *`location(newLocation)`* to the user. The *`newLocation`* parameter is an object that implements the `Location` interface.

Once the `Location.User` receives the message it triggers a change. Since the user in the service layer is dependent of the `Location.User` it gets an update which, in turn, triggers a change that is captured by the `ServiceEnvironment`. When the environment gets this notification it calculates (by interacting with the user and the available service areas) which service areas the user left (if

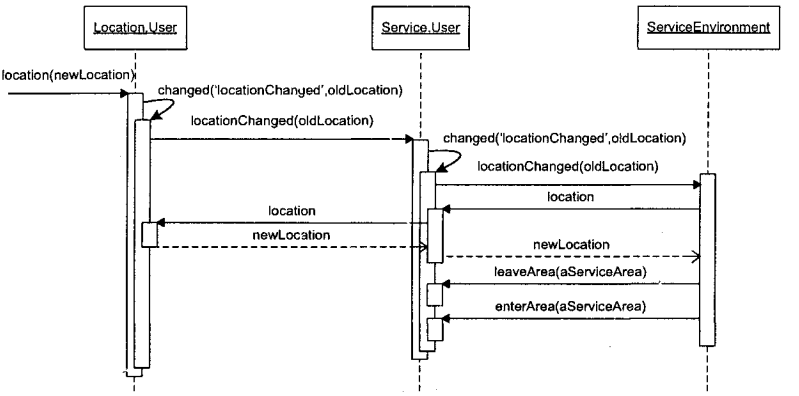


Figure 6. A three layer architecture for Location-Aware Services

any) and which ones he entered. After the message `leaveArea` (or `enterArea`) is sent to the user, he will end with old services removed (or new added).

4.7 Adding other Context Models

In this section we briefly describe how we can add time constraints to our application using the same philosophy described in the paper. Suppose that we want to specify that a service is available at a certain area in a particular period of time. Following with the campus example, we would expect the course material service to be available at the time the course is being held; once the lecture is over, the material shouldn't be accessible as a room service. To implement this mechanism we must first add the notion of time and time events to the context abstractions; in order to do so we add a Time package containing the class `Timer`. The `Timer` class is a Singleton (Gamma et al., 1995) that has two main responsibilities: it can be queried for the current time and it can be configured to send time events in predefined moments.

Once this package is added to the application, we need to configure our services to have a time constraint: the service will only be activated in a predefined period of time. At first glance, the implementation of this constraint seems to be straightforward: as we have seen before, when the user enters a room, he triggers a change that ends with the user asking for the services available in a service area. When a service area is asked for its available services, it checks the service's constraints; a service will be available if and only if the time constraint is satisfied (this can be verified by asking the timer for the current time). Now, suppose that the user enters the room before the course starts; since the time constraint is not satisfied, the course's material is not presented to the user. After a couple of minutes the course begins, but since the user is

still inside the room (i.e. he hasn't left the room and re-entered it again) he doesn't have the course material service. The problem in this case is that, so far, service changes are only triggered by location changes, while time changes should also affect the services available for a user. To solve this problem, we need to be able to configure time events associated with time constraints: when a service is accessible during a specific period of time, time events should be generated at the beginning and at the end of the period, so that the services available for a user are re-evaluated. The events generated by the timer, are captured by the dependency mechanism and dispatched to the environment, which in turn asks the user object to analyze again the services provided by the service area that has just changed.

5. AN ARCHETYPICAL IMPLEMENTATION

We have built a proof of concept of our architectural framework using a pure object oriented environment (VisualWorks Smalltalk) that supports dependency mechanisms and reflection, and where truly transparent distribution can be implemented. To achieve distributed objects collaboration in a transparent way we used the Opentalk framework, which we adapted to support PDAs sockets; we also extended the framework to perform object migration from one device to another. We used HP iPaq 2210 PDAs as user devices; user's location sensing was performed using infrared beacons and we are now adapting the sensing mechanism to work with bluetooth signals.

Our design prototype is not conceived to work on a client-server style, but mainly on a fully distributed environment shared between different devices. This approach promotes an environment where we can find different kinds of PDAs and desktop machines working together in a transparent way. We have filled our expectations so far, since we are interacting with wireless PDAs and wired PCs without any trouble; we also have upgraded our PDA hardware to HP iPaq hx2750 without even noticing it.

6. RELATED WORK

We found our model of context to be quite similar with the one presented by Dourish (Dourish, 2004). While in most approaches, context is viewed as a collection of data that can be specified at design time and whose structure is supposed to remain unaltered during the lifetime of the application, Dourish proposes a phenomenological view of context. In this approach, context is considered as an emergent of the relationship and interaction of the entities involved in a given situation. Similarly, in our approach, context is not treated as data on which rules or functions act, but it is the result of the interaction between objects, each one modeling a given context concern. In addition, we

do not assume a fixed context shape, and even allow run-time changes on the context model.

From an architectural point of view, our work can be rooted to the Context Toolkit (Dey, 2001) which is one of the first approaches in which sensing, interpretation and use of context information is clearly decoupled. We obviously share this philosophy though pretend to take it one step further, attacking inner application concerns. Hydrogen (Hofer et al., 2003) introduces some improvements to the capture, interpretation and delivery of context information with respect to the seminal work of the Context Toolkit. However, both fail to provide cues about how application objects should be structured to seamlessly interact with the sensing layers. Our approach proposes a clear separation of concerns between those object features that are "context-free", those that involve context-sensitive information (like location and time) and the context-aware services. By placing these aspects in separated layers, we obtain modular applications in which modifications in one layer barely impact in others. From an architectural point of view, our work has been inspired in (Beck and Johnson, 1994): the sum of our micro-architectural decisions (such as using dependencies or decorators) also generate a strong, evolvable architecture.

In the Java Context Aware Framework (Bardram, 2005), a Java-based framework is presented for building context-aware applications. Even though the framework presents a behavior oriented structure, it still models context in a traditional way (by means of context and context items) and makes an explicit separation between the entities and their context (in fact, entities explicitly know their context). In our proposal, we think of context as extending the base application behavior instead of viewing context as data to be acted upon. Since the layers are built on top of the application model, there is no need to change the core of the system in order to make it context-aware.

To summarize, in our approach we see *context aspects* as active objects and the context itself as an emerging property of their interaction. To achieve independence between the contexts aspects and the sensing mechanisms we placed a layer between them, so that changes in one model does not affect the other. At the architectural level, and thanks to the increasing power of the mobile devices, we decided to work with distributed objects instead of using a client-server architecture. In this way the applications running on the PDAs are responsible of handling the services of each user and can provide more advanced services than the ones provided by web pages, avoiding at the same time the scalability problems associated with concentrating all the processing in a single server. Lastly, in order to be isolated from lower level details, we decided to implement our framework on a pure object oriented environment as Smalltalk.

7. CONCLUDING REMARKS AND FURTHER WORK

We have presented a new approach for designing location aware services and described how to enhance existing applications with new context-aware behaviors. By using a dependency mechanism to connect locations, services and application objects we have been able to avoid cluttering the application with rules. We have also improved separation of different design concerns, such as applicative, spatial, temporal, sensing, etc. Additionally, we showed how to achieve a finer granularity of design concerns with respect to existing approaches.

Our view represents a step forward with respect to existing approaches in which context information is treated as plain data that has to be queried to provide adaptive behavior. We briefly described a prototype system that we are using as a proof of concept for building context-aware services.

We are now working on the definition of a composite location system that allows symbolic and geometric location models to co-exist seamlessly. We are also planning to enhance the simple dependency mechanism to a complete event-based approach, delegating specific behavior to events and improving at the same time the framework's reusability. We are additionally researching on interface aspects to improve presentation of large number of services.

References

Abowd, G. D. (1999). Software engineering issues for ubiquitous computing. In *ICSE '99: Proceedings of the 21st international conference on Software engineering*, pages 75–84, Los Alamitos, CA, USA. IEEE Computer Society Press.

Bardram, J. E. (2005). The java context awareness framework (jcaf) - a service infrastructure and programming framework for context-aware applications. In *Pervasive*, pages 98–115.

Beck, K. and Johnson, R. E. (1994). Patterns generate architectures. In *ECOOP*, pages 139–149.

Dey, A. (2001). *Providing Architectural Support for Building Context-Aware Applications*. PhD thesis, Georgia Institute of Technology.

Dourish, P. (2004). What we talk about when we talk about context. *Personal and Ubiquitous Computing*, 8(1):19–30.

Gamma, E., Helm, R., and Johnson, R. (1995). *Design Patterns. Elements of Reusable Object-Oriented Software*. Addison-Wesley Professional Computing Series. Addison-Wesley.

Hofer, T., Schwinger, W., Pichler, M., Leonhartsberger, G., Altmann, J., and Retschitzegger, W. (2003). Context-awareness on mobile devices - the hydrogen approach. In *HICSS*, page 292.

Leonhardt, U. (1998). *Supporting Location-Awareness in Open Distributed Systems*. PhD thesis, Dept. of Computing, Imperial College.

Rossi, G., Gordillo, S., and Lyardet, F. (2005). Design patterns for context aware adaptation, Workshop on Context-aware Adaptation and Personalization for the Mobile Internet.

Salber, D., Dey, A. K., and Abowd, G. D. (1999). The context toolkit: Aiding the development of context-enabled applications. In *CHI*, pages 434–441.

Sousa, J. P. and Garlan, D. (2002). Aura: an architectural framework for user mobility in ubiquitous computing environments. In *WICSA*, pages 29–43.

DEPLOYMENT AND USE OF MOBILE INFORMATION SYSTEMS

A case study of police work

Alistair Norman and David Allen
Leeds University Business School

Abstract: The paper presents the results of the investigation of the implementation of mobile technologies in an under researched area: the police. Five key themes of investigation with relation to mobile information and communications technologies were identified in the research: changes in work procedures, changes in the organisational capability, changes in relationships, effectiveness of equipment and effectiveness of infrastructure. These themes provide a framework for analysis of the police context, and one which could perhaps be extended to other contexts in the public safety / service arena.

Keywords: Implementation of mobile technology, case study

1. INTRODUCTION

This paper reports the initial impacts of the introduction of mobile information and communications technologies into a UK Police force. The paper aims to describe the officers' reactions and to develop a first stage model of the areas of attention in the process of introducing such technologies. Despite the growth in use of mobile information and communications technologies (MICT), the huge potential opportunities they offer and the challenges that they pose, there has been an emphasis in the literature to date on the stationary applications of computing (Kristofferson and Ljungberg, 1999a; Weilenmann, 2001). This relative neglect of the mobile technologies is not, however, unrecognised and Weilenmann comments that this lack of research into MICT is not the only gap and notes that other existing technologies for mobile working, such as the VHF radio and its use by organisations, are poorly researched. Whilst the lack has been

addressed to some extent by recent work such as that of Nulden (2003) and Pica et al (2004) which have looked at the specific police context and that of Sawyer and Tapia (2003) which has looked at the public safety context more broadly there is still a recognition within this work, and in parallel work such as Vaast and Walsham (2005) that we actually know relatively little about the changes in work practice which result from the introduction of information and communication technologies into work settings. Vaast and Walsham comment that we can posit confidently that there is an effect from the introduction of ICT into work settings but that we have little understanding of exactly how work practices change with ICT use and how the interrelationship of technological artefact and technology-in-use develop. Indeed, while mobile environments are becoming more common writers such as Pica et al (2004), Yoo and Lyytinen(2003), Bellotti and Bly (1996),and Kristofferson and Ljungberg (1999 b) have identified that such environments are qualitatively different from fixed computing and that they both pose new variants of old problems and raise new problems for those seeking to make best use of the technologies. Yoo and Lyytinen, for example, comment that mobile systems are unlikely to prove any easier to calculate returns from than fixed systems and suggest that as a result of their distributed nature and 'multiple levels in multiple contexts' they may well prove to more intractable than more traditional systems when it comes to showing a positive return on investment. Weilenmann (2003) and Green (2002) suggest that mobile information and communications technologies (MICT) can lead to a potentially detrimental blurring of boundaries between work and other activities and Kristofferson (1999 c) comments that current platforms 'do not realise the full potential of mobile computing as mobile work and IT use differ significantly from other settings' This concern is reflected in the finding that the use of the technology is currently mainly a white collar phenomenon (Brodie and Perry, 2001) and, on a linked point Nishibe and Waki (1998), dealing with MICT support for academic conferences, note that the majority of applications are small scale. They report on a range of contexts of use but note that the challenge may not be to get small scale trials to work, but to scale this up to commercially useful levels.

This research project focused on the use and deployment of MICTs in the particular context of policing in the UK. The UK Government has recently invested over four billion pounds in a mobile communications network for UK police forces. The continuing cost of this infrastructure is in excess of 200 million pounds per year of central government funds and an unspecified amount of money at local police force level. This network provides voice and limited data communication capability. In addition to this activity many local forces are heavily investing in other mobile information and communication technologies in an attempt to provide further functionality.

Despite the significant changes anticipated by the development of this technology, there has been little academic interest in this area. This paper analyses one of the earliest and largest deployments of mobile information and communication technologies within the policing environment. Until relatively recently only a few articles explicitly mentioned the application of such technologies to policing. Colton (1979) for example identifies the potential of ICT for policing and Nunn (2001) noted that Police forces tend to be major users of ICT and posited that the reasons for this are the normal business drivers of improving efficiency and effectiveness, commenting, 'At the simplest level IT is added in order to improve things'. This lack of research is despite the fact that police officers in most developed countries operate in an information rich environment (Nulden, 2000). They have access to sophisticated databases (Hauck and Chen, 2002) and the facility to call on colleagues via email, mobile phones, radio contact and in personal meetings. Recently there has been an upsurge of interest in the area of public safety as a whole (Sawyer and Tapia (2003 for example) and in the police specifically. Pica et al (2004) outlined the differences in police roles and the issues of passive vs. active and structured vs. unstructured information use in police work. Nulden (2003) has addressed the difference between police and policing and has suggested a framework to determine whether mobile technologies provide advantage for officers. A common theme in all of this work is a recognition that the devices themselves are not the focus of analysis – it is the use of the devices in context which is both more valuable and more complex. This has been neatly summed up by Sorensen (2003) who commented that technology use is about more than technology and clearly stated the need to 'take account of the actualities of human interaction' In evaluating the effectiveness of the ICT investment Yoo and Lyytinen (2003) highlight the problems of measuring impact in ubiquitous computing environments and Nunn (2001) suggests that 'the jury is still very much out' on the value of ICT. He identifies specifically that the cost associated with ICT can mean that there are fewer officers to deliver on the ground services and higher average costs for technical personnel. This may well be seen to conflict with current UK political commitments to high visibility policing with 'more officers on the beat'.

The rest of this paper is in four sections, the next section of this paper briefly discusses the nature of mobile work and mobility. This is followed by a section briefly describing the research methodology and methods. The findings from the research are presented in the next section and the final section provides some conclusions and areas for future consideration.

2. THE NATURE OF MOBILITY

In discussing mobile work and the technology available to support it there tends to be an assumption that we know what is meant by 'mobility'. Where attention has been paid to this issue it has tended to concentrate on geographic mobility to the exclusion of other areas. An embryonic literature is, however, emerging which has started to examine the nature of mobility and the impacts on the way we work more broadly. The following review of the literature identifies and briefly reviews three core areas. The first reviews the core analytical frameworks that have been put forward as analytical frameworks for understanding mobility. The second group of literatures relates to the use of mobile technologies within the context of the workplace. The third literature relates to the organisational benefits that are related to the deployment of the mobile information and communication technologies within the workplace.

Kakihara and Sorensen (2001, 2002a, 2002b, 2002c) Green (2002) and Kristofferson and Ljungberg (1996), in particular, have started to extend the notion of mobility. Kakihara and Sorensen have written extensively on the nature of mobility and argue that we need to examine more than the location where people transact work if we are to get a true picture of the nature of the work and the impact and utility of technology in that situation. Kakihara and Sorensen (2001, 2002 a) presented an initial framework which has been extended by their later work in a number of areas. They argue that people can be mobile in different senses. The first is geographic or spatial mobility which is much discussed in the literature but only really in terms of the movement of people. Kakihara and Sorensen (2001) argue that this is a restricted view and that we need to look at a number of additional facets of this physical movement of people including, the mobility of objects, symbols and space as a result of symbolic travel on the Web. The second facet of mobility is temporal and it has been argued that the availability of mobile technologies has altered the nature of temporal constraints on work. At the crudest level a mobile telephone allows you to call a colleague on the other side of the world and leave a message for them in the middle of their night and your day. On a slightly more complex level they also argue that this is in some senses a commodification of time and lead in to a discussion of monochronicity and polychronicity; polychronicity being the divergent use of time rather than adherence to a pre planned order – monochronicity - and they note that MICT can increase both. This apparently contradictory position is reconcilable; MICT promotes polychronicity in that it can allow us to work on a number of tasks at the same time or outside of the normal time frame within which we would deal with them, such as taking a work call during a social event; it also allows us to summon up information or resources to allow us to continue with a task which, without MICT, would

have had to be shelved until we returned to the office, or a fixed terminal. This notion of time flexibility and the compression of activity into shorter 'soundbites' is a facet of mobility which is also noted by Green (2002). The third facet of mobility which Kakihara and Sorensen (2001) discuss is that of contextuality. ICT allows people to be free of many contextual constraints; so, for example, it is possible to make a social call from a business context or check a stock price on a PDA at a birthday party, but it also imposes the risk that others will not be sensitive to the context that you are in when they try to interact with you (Fitzmaurice, 2000). Kakihara and Sorensen (2002 b) identify that while devices do allow some limited declaration of context to others there are dimensions of the use of MICT which can be detrimental.

In order to understand the nature of mobile interaction Kakihara and Sorensen (2001, 2002 b) suggest that it can be considered in terms of regions, networks and fluids. Regions, with fixed boundaries and a sense of enclosure are a metaphor for the pre ICT organisation, networks are a metaphor which work well for physically connected ICT and the fluid metaphor is the one they use to discuss and analyse MICT. In discussing this Kakihara and Sorensen, (2002 b) give an example of a delivery firm using a range of mobile technologies to manage a complex environment and act proactively to meet the constantly changing needs of their customers and (2002 a) of consultants working in a fluid way across a range of organisations balancing a range of tasks, locations and demands; a way of working which they characterise as 'post modern professionals'. They argue the need for an organising paradigm for mobile working and the use of technology within this mobile working environment to combat the dangers of detachment from reality and the increasing blurring of the boundaries between work and social life.

Kristofferson and Ljungberg (1996, 1999 c) identify three key modes of mobility which they argue can be used to start to understand the way in which people use mobile technologies. These three modes are 'travelling, visiting and wandering' and are illustrated with examples of the types of activities which are undertaken in these ways and the technologies which are needed to support them; so the traveller requires the ability to move information but does not need to be able to manipulate that data en route, the visitor needs to be able to gain access to information from a range of fixed locations, and the wanderer needs to be able to access the information on the move probably without fixed connections. An illustration of the wanderer is provided by Gallis (2000) in a medical context where staff cannot guarantee to be able to access fixed technology and so need to be able to make full use of the potentials of MICT.

In a more recent article Pica and Kakihara (2003) suggest that the existing concept of mobility is a somewhat impoverished one in many discussions and they suggest that concepts of a dualism of views of mobility

centred around views of stability and of fluidity separately need to be replaced by a duality 'studying both fluidity and stability in contemporary society and work organizations and understanding their mutual influences' This paper highlights the need for further investigative work producing 'rich and contextualised data' which integrates a micro (interactional) viewpoint with a macro (organizational) viewpoint. Sorensen (2003) also highlights the need for a richer definition of mobility and comments that for many professionals and knowledge workers MICT means that they have 'everywhere to go and nowhere to hide'.

The police context is one where the officers have been mobile workers for a long time and have had the support of communication tools and management tools to allow them to undertake a role that is often reactive, highly mobile and dependent on good information for the effective and defensible performance of the role. MICT has changed the landscape of the police organisations by changing the nature of the mobility of police officers, by allowing them to relate to peers, managers and the public in different ways and by providing ways to both automate and increase the efficiency of existing processes as well as adding the ability in some areas to police their communities in ways which were previously not available to them.

3. RESEARCH METHODOLOGY

The approach taken in this research was interpretive and qualitative. The analytical frame used was Grounded Theory. This approach has been extensively used within the Information Systems Research Community (Howcroft and Hughes, 1999). Grounded theory requires a researcher to examine the data collected and build up 'theories of process, sequence and change pertaining to organisations, positions and social interaction' (Glaser and Strauss, 1967) and is inductive, contextual and processual (Orlikowski 1993) requiring an iterative approach to the collection and analysis of data and constant comparison across evidence to control the conceptual level and scope of the analysis. A hermeneutic approach is implicit in the analysis of data collected in an interpretive paradigm and elaborated using grounded theory, it is also applied as a specific analytical technique in this research project and this was facilitated by the use of Atlas ti, which is discussed below. The hermeneutic circle is predicated on a movement 'from the whole to the part and back to the whole' (Gadamer, 1976) which is intended to 'try to make sense of the whole, and the relationship between people, the organisation and the information technology' (Myers, 1997).

Two groups of mobile users were interviewed, the first are a group of Criminal Investigation Department (CID) officers who have had laptops and

mobile Global System for Mobiles (GSM) connections giving them remote access to all network facilities on a laptop Personal Computer (PC) for about eighteen months. The second group of staff are Scene Of Crime Officers (SOCO) who have been provided with similar equipment together with the specialist software and hardware required to operate a specialist role. This specialist software package was not operational at the time of the interviews with these officers and they were only just getting used to having mobile access to the main applications discussed in the introduction. The sample provided access, therefore, to users with experience and to novice users. It also provided a coverage of two quite different roles – CID officers are often quite heavily office based and SOCOs spend much of their time out of the office in what they describe as a nomadic life moving from crime scene to crime scene; their demands of technology are also quite different – CID officers need access to office type facilities and to information sources such as databases while SOCOs need to be able to access specialist software. Fourteen interviews were conducted with individual officers. Interviews lasted 60-90 minutes. The interviews were guided by the use of a semi-structured questionnaire which was refined after each theory generating iteration.

Across this period observations were also undertaken, in the incident handling centre (IHC), in the CID offices and on site with SOCOs. Interviews were also conducted with two key senior staff – the officer managing the Mobile Technology and Knowledge Programme and the Senior Management sponsor. These observations and interviews were scene setting and were used to enable the writer to have a better understanding of the context in which officers and the technology operated.

Interview data was transcribed verbatim from the tapes made at the time of the interview. This was done using a voice recognition package to dictate the interviews into a word processor and the files were then sense-checked and compared to the original tapes for accuracy. This data was analysed using the computer programme, Atlas ti. The package makes use of the techniques and approaches of grounded theory and takes a 'hermeneutic unit' of text or graphic resources as its base. In this study all resources used were in the form of .txt files transcribed from the interviews with police officers. The process is, therefore one where the 'two analytical techniques of theoretical sampling and constant comparison ... are the means by which Grounded Theory proceeds' (Hughes and Howcroft, 2000). In this study the coding process was started after the first round of interviews and built up as further interviews were conducted. In all, interviews were conducted in three rounds across a period of six weeks. Codes were iteratively built up, and quotations attached to these codes. These codes were grouped within the package and memos developed to provide the basis for areas of analysis and investigation within the text data. These codes and attached quotations were

then organised into five categories and a total of fourteen sub-categories within which to examine the use of MICT.

4. DISCUSSION AND RESULTS

The analysis of the interviews with officers produced five main categories under which their comments on MICT could be grouped. Two of these categories – equipment and infrastructure – are in the nature of 'hygiene factors' in that a failure to ensure that they support effective use will reduce the level of use to a point where it is not useful to observe changes in the other three categories. Each of these categories has a number of sub categories associated with it and these have been used to build up an initial framework for examining the impact of MICT on officers and on the organisations within which they work. Table one below summarises the categories and sub categories derived from analysis of interview data.

Table 1: Categories and sub categories derived from analysis of interview data

Category	**Sub categories**
Changes to work procedures	Changes to officers ways of working Changes to boundaries and controls on working
New capabilities as a result of the adoption of MICT	Working faster or better at a level above that of incremental or evolutionary improvement Working in new or radically different ways
Relationships	With colleagues With supervisors With others With the public
Equipment	Concerns about loss or fragility of equipment Issues relating to ease of use Issues relating to connectivity
Infrastructure	Ways of working Support issues Future potential

In this section of the paper each of these categories are discussed in turn. The final part of this section briefly discusses a framework which provides a mechanism for understanding the relationships between the different categories.

4.1 Changes to Work Procedures

This category is concerned with the way in which the work is done and the ability of the users to use the technology to achieve their work goals. The analysis provided two sub-categories; changes to the officers own ways of working, and changes to boundaries and controls on working.

Within the sub category of 'changes to officers own ways of working' officers identified that the use of MICT allowed greater control over their own time, including the use of slow time (this referred to time which wasn't spent responding to an incident or dealing with a member of the public) , to complete work. Officers identified that the technology allowed them to reschedule tasks as well as not having to wait for information to be relayed via an information intermediary, as one officer commented;

'I definitely multi-task but it makes it a lot easier when you have access directly to information – you can look at your workload and work more jobs and give them a decent amount of time whereas before you would be waiting for something and then forget what you had been working on before' .

The feedback from officers in this sub category officers picked up on a strong theme in the literature (Kakihara and Sorensen, 2001, 2002a, 2002,b; Green, 2002; Lindroth et al., 2001, Edvardsson and Bergqvist, 2002, Juhlin, 2001) which identifies that the use of MICT can allow users to control and manage time in new and potentially more effective ways, and specific reference was made to the use of the technology in ways such that both monochronicity and polychronicity can be supported (Kakihara and Sorensen, 2002a) giving officers the choice of multitasking where appropriate and also allowing them to follow a job through where this is appropriate. The ability to spend more time in the operational situation was also noted as a positive benefit by many officers and reflects the findings of studies in sales forces (Krogstie, 2003; Watad and DiSanzo,2000) and in maintenance settings (Pakanen, 2001; Nielsen, 2002, Wiberg, 2002).

Officers also noted the potential to make use of the technology in the form of more efficient scheduling of their tasks; this is especially important for officers like SOCOs who can criss-cross an area visiting scenes of crime, rather than being able to plan an efficient schedule, due to lack of information. SOCOs see this as a key advantage of mobile technology and one summed this up 'so in many cases the kit allows access to better scheduling and maybe allows you to pick up something which has not come over the PR'. Again this is a theme which recurs in the broader literature with scheduling efficiencies being identified in a number of studies (Pica et al, 2003; Juhlin 2000; Jipping et al, 2001)

MICT was seen as a complement to existing communication channels rather than a replacement for them. This provides officers with the chance to take transactions off the air and onto the mobile technology, freeing up

airtime and also allowing them to be more thorough. One officer said 'for us I think it's going to be that the information is going to be there quicker and we can do more delving than we would dare to do over the radio'. All of the officers who commented on this issue saw this as being a positive development. The use of the equipment to take communication off existing communication channels also reflects an identifiable theme in the literature (Kakihara and Sorensen, 2001; Abowd and Mynatt, 1997).

MICT was also seen as being a key way to access decision support materials, allowing officers to make better decisions without having to make reference to printed materials or to colleagues. This was seen as a way to be more certain in dealings with unfamiliar situations and one patrol officer commented;

> *'normally we would just crash ahead and do things – or hold off and think that we would come back in here [the police station] and look them up. Perhaps we could just go ahead and do them on the street without having to lug a couple of big fat books around.'*

Other effects identified included the issue of reduction of equipment, analogous to the 'reduction of the backpack' identified by users in educational settings (Soloway et al., 1999). This was reinforced by other officers who commented that they took a bulk of material with them on patrols (of which they used very little on an average patrol) which could be reduced by placing reference materials onto mobile technology. The comment was made by officers, with some force, that the equipment is useful only when it provides business benefit and Nulden (2003) identified resistance to the use of MICT in a police setting as the device per se was not seen to be advantageous to the officers concerned – in addition he suggested a framework of functionality, properties, modalities of use and mandate for examining the use of the technology-in-use. In addition, one potential for the future identified here was that the mobile technology could eventually replace the pocketbook, an illustration of the potential of such a replacement is the use, identified by five officers, of MICT to allow officers to update CIS files with intelligence updates at the time that the information comes to the attention of the officer rather than having to wait until the officer has access to a shared terminal at the police station. This is linked to the ability of the technology to update information from the field which is a theme in the literature (Guerlain et al., 1999; Watad and DiSanzo, 2000).

The risk of losing the nuances of personal communication has been identified at a macro level (Green, 2002) and at a micro level (Luff and Heath, 1998) and Green (2002) also notes the issue of resentment at being 'kept tabs on' remotely as does Weilenmann (2001). The 'blurring of the boundaries between social and work life' noted by Green is also reflected by

the officers in these interviews and this is an issue discussed extensively by Kakihara and Sorensen (2001, 2002b, 2002c).

In the sub category of ' changes to boundaries and controls on working' the main issue was the existence of comprehensive log files allowing the use of MICT to be audited. Officers perceived that this provided them with reassurance that there are external logs to confirm the probity and accuracy of their actions. Others, however, were concerned that the existence of such files would be used in order to observe productivity and working patterns and some police officers interviewed expressed the concern that the technology could be used as a 'big brother' form of control of officers working time. This concern was vague in content but attracted comment from a number of officers, one of whom said '[I] would be concerned that we were being looked at or monitored, and I think that is bound to happen, but I wonder in what form it will be and what exactly they will be looking at?'.

There were also areas which were raised by the officers which are either ephemerally reflected in the literature or are not reflected at all. Most of these issues could be regarded as police-specific. There was a strongly expressed concern that the nature of the technology could expose officers to a range of risks to health and safety – this is not an area dealt with to any extent in the literature beyond the issues related to the problems of using small interfaces to get information into and out of mobile devices and yet was of major concern to officers in the operational situation who identified the problems of portability (or lack of it with laptops), inattention or distraction and the failure to call in location endangering officer safety. Officers also noted the dangers of using the devices in quicktime (response incidents as discussed above) incidents when attention needs to be paid to the unfolding events rather than the information, which would be more appropriately handled by information intermediaries.

4.2 Changes in Capability

This category is concerned with the speed, thoroughness and accuracy of work done by officers. In the analysis two main sub-categories emerged within the overall category. The first is concerned with the extent to which work would be affected by being able to do things faster or better than they are done without the use of MICT – to the extent that they add a significant new capability to the manner in which an officer can work. An example could be the use of MICT to check a selection of car registrations at 2.00 am on Sunday morning, something which would not be possible over the radio at that time of the night. The second is concerned with the extent to which MICT could enable new or different work practices to those currently in place and an example in this area could be the ability of the mobile

application of the Home Office Large Major Enquiry System (HOLMES 2) to integrate information into a database at the point of entry rather than having to wait for up to two weeks for this information to be dealt with in the current 'normal course of events'

With regard to being able to do the job 'faster or better' officers identified a number of strengths / potential strengths. Overwhelmingly officers felt that the availability of information via MICT would allow them to do their jobs better. This positive attitude was based on the expectation that the MICT facility would provide them with access to data faster and this is reflected in studies by Lindgren and Wiberg (2000), Lindroth, (2000) and Nielsen (2001). Timeliness is an issue of importance to the policing context and this is discussed both by Nielsen (2001) and by Guerlain et al. (1999) who identify the role of MICT in placing information where it is needed, when it is needed and in the form that it is needed.

Almost all the comments made about quality of information supported the view that mobile technology would mean that there was more use of information and that officers could be more thorough in the performance of their duties. This was primarily due to the availability of more information – and this was seen in part as a function of being able to do more checks than would be the case when working via the IHC. As one officer commented there is a value judgement to be made as to whether it is 'worth' doing a check and this value judgement is different depending on the demands on the IHC at the time; this officer said that 'there are a lot of people who we stop who we know will probably come back alright, so we don't bother to do the check. With this we could take a minute and so we would just do it.' Another officer felt that the technology (a laptop in this case) 'allows us to do more checks before we walk in the door, and that increases the chances of having insights to share with other people'. Many positive comments were concerned with increasing the speed with which information could be supplied to officers in the field, although one officer noted that there was a potential for overload in the availability of information 'I can do more and faster; so in some respects it is wearing me out because I'm actually doing a massive amount more'. Officers also noted that MICT should speed up information supplied via the radio systems as a result of the reduction in airtime congestion. At present the 'airtime does get really busy at times' so transferring information requests to MICT would allow those officers needing to use broadcast systems to be able to do so with less delay and this freeing of the radio for urgent jobs was a specific strength noted by seven officers. In conjunction with this increase in speed was a suggestion, by over three quarters of officers interviewed, that the accuracy of information received would be higher over MICT than via the IHC. This was expected to be a result of officers entering data queries themselves rather than via

information intermediaries who may mishear or misspell a name or reference and chasing things further. A typical comment was that;

> *' If you have used CIS you know there are ways of searching and the search engines on the system; you know you can search in so many ways but if you go via control they are just too busy and they often come back with "No results here." And you do your checks yourself when you get back and you think "Christ, this has come up, that has come up"- and that would have affected the way you dealt with the job.'*

It was also recognized by officers that the IHC staff would welcome the reduction in traffic, allowing them to concentrate on dealing with major or 'quicktime' incidents and this was an aspect of general ' reducing the call on others' which officers saw as a strength of MICT. The use of the equipment to take communication off existing communication channels has already been noted above (Kakihara and Sorensen, 2001; Abowd and Mynatt, 1997).

The quality of information retrieved was also seen as being likely to improve and this is an issue highlighted in the literature by Wiberg (2000) although it has not been investigated in any depth. In the case of the police context the increase in accuracy was expected to come from the ability of the officers on the ground to delve further using MICT than they would using the current PR systems and from the fact that they were inputting information and removing a link from the current supply chain for information and, with this link removing a source of error (poor sound quality, mis-spelling and lack of time to deal with queries being cited). Officers perceived the technology as making them more independent and this is again, a theme which has been identified in the literature to date in studies such as those by Juhlin and Normark (2000), Laurier and Philo (1998) and Watad and DiSanzo (2000) although fears were also expressed that this could lead to isolation as identified by Pakanen (2001).

The ability to access forms and documents remotely was identified as a strength by many officers, especially when taken with the ability to use slow time. One officer commented 'you can just park up, especially on nights when it is a bit quiet, and bang it out [routine paperwork] and it is done'. The use of efficiency enhancing applications, and specifically spreadsheets, was mentioned in comments from five officers as being a strength of MICT. One officer commented that the efficiency gains meant that she would tackle jobs which would not have been attempted before and she depended heavily on the facility, saying 'I use the spreadsheet for everything and it is fantastic for that – so, whereas in the old days I would have had to write up 800 exhibits onto a witness statement now I just copy and paste off the sheet and it is done.'. A potentially important issue was also raised under this sub-

category with regard to the general quality of presentation of work with officers commenting that the use of IT allowed them to present work more professionally and improved the access to information, one officer noted that this was linked to facility with keyboards more generally 'their [new entrants] typing skills are phenomenal...so whereas before you would get a handover package and it was scribbled and you couldn't read it, if you get a tight and clear, legible document that sets out everything you need'.

There were, however, also some significant concerns relating to this sub-category of changes in capability. Concerns with relation to speed were expressed by officers interviewed in three main areas. The first was that officers would not update information at the time when the information came to their notice but might rely on imperfect recollection to update intelligence files at a later stage; a typical comment from an officer was that 'we might have an awful lot of knowledge in our head, but we are thinking about the next job and we just want to move on and do it. And that could come back to bite us later if something goes wrong for any reason whatsoever'. The second issue, although a minority view, was still significant and was a concern that the availability of e-mail via MICT would be a starting point for information overload for the officers concerned. A third concern was around the potentially slow speed of supply of information using mobile browsers and wireless connection and three officers expressed fears that the connections would be too slow for meaningful work. One senior officer commented on this topic that 'I currently use my mobile with my laptop for mobile data, but the bandwidth is so small that apart from very quickly looking at ICAD it is hardly worth using at all'. Officers also noted that the ability of the technology is currently limited by the lack of pervasive connections allowing databases and information to move seamlessly across networks and devices, and this is an area that officers saw as being aspirational for the use of MICT. The literature on pervasive (Fails and Olsen, 2002) or 'everyday' computing (Abowd and Mynatt, 2000) identifies this but offers no immediate solutions.

A final concern is that there will be a level of loss of incidental knowledge as a result of the transfer to a more personal information environment. Currently much police information is either available to all as a result of being on the force networks or is explicitly given to all members of a team in a briefing or by virtue of coming over the voice radio system. Officers have a clear idea of what their team are doing and this situational awareness is cited as a strength of the current systems by some officers. MICT allows information to be sent to an individual officer in accordance with their role or their availability and so a lot of information which was formerly public as it was passed to the individual(s) for who it was meant, will now go directly to them. This area was not a concern for all officers but was one of the expressed fears that the police put forward at the start of the

study and is reflected in the literature by Wiberg (2000). A significant minority of officers interviewed also raised the issue that their skills in inputting information to MICT were such that the accuracy of the information retrieved could be compromised as a result of their lack of IT skills.

4.3 Relationships

This category is concerned with the impact that the use of MICT will or may have on the relationships which officers have as a part of their working lives. The category attracted significant comment and most of these comments were positive about the use of MICT, identifying strengths or potential strengths and, where there are concerns about weaknesses or potential weaknesses, they tended to be expressed less forcefully than in other categories with only a few codes showing solely concerns. In the analysis three key sub-categories emerged; these were, (1) relationships with colleagues, (2) relationships with supervisors, and (3) relationships with others.

Taking the sub-category of relationship with colleagues first, it was noted that IT and MICT in particular has the potential to make people more independent, able to do the job 'without relying on someone else in another room' as one officer put it. The ability of users to act more independently with greater control of time and their manner of working is one which is reflected in a number of studies including Wiberg (2000) and Watad and DiSanzo (2000). It was also noted that this is not far from making someone more isolated. The analysis revealed that there were a significant number of comments praising potential independence and an equal number expressing fears of potential isolation. There were also areas of weakness or potential weakness which have been identified in the literature and which were also of concern to the officers interviewed. One of the main issues was the potential isolation which some officers could experience. Many officers, especially SOCOs and Community officers work independently and there was an expressed fear that their 'road warrior' lifestyle would be made even more lonely by the use of MICT. This risk of isolation has also been identified by a number of writers in a range of settings (Krogstie, 2003; Weilenmann, 2003; Green, 2002; Wiberg, 2000). One of the issues which had been identified at the initial stage of investigation was that potentially the use of MICT could reduce camaraderie and team ethos. This does not seem to be a concern at a significant level for the officers interviewed and all of the fourteen comments on this code said that they did not foresee this as a problem, a typical comment being ' in the office we will still chat amongst ourselves and there is a good camaraderie. We still have briefings and we still chat'. Officers felt that camaraderie came from team activities and

briefings far more than it did from the open information environment currently in place. The use of MICT is not expected to damage team ethos and camaraderie (although this was a fear for the managers at the start of the study) and both Okoli et al. (2002) and Gaines and Shaw (1994) note that a positive effect on camaraderie has been a benefit of the use of MICT in their respective studies

All of the comments dealing with the topic noted that the ability to reduce their call on others for information could significantly improve their working relationship with those others, and this has been noted under other categories above. Generally officers have a positive view both of MICT and of the effect they believe it will have on their relationship with their peers, although some noted that there was a danger of being perceived as a 'nerd' either by colleagues or by the public (akin to the perception of the equipment as a Gadget – Lindroth et al. (2000)).

As far as the relationship with supervisors is concerned most officers felt that the introduction of MICT was unlikely to change their relationship with supervisors and only about a few of the comments on this topic expressed fears about the role of the supervisor. This had been a concern expressed by management at the initial stage of discussions with the police force but, overwhelmingly, officers felt that a good supervisor would be aware of the needs and strengths of their team through whatever communication channels were available. As one officer commented *'A good supervisor is a good supervisor – this is just another tool. The difference is between those who understand that "supervise" is a verb and means you do something, and others who believe it is a title and things happen round you'.*

The relationship with the public was felt to be one which would benefit most from the introduction of MICT with officers overwhelmingly seeing this as something which can present a positive image to the public. This positive image is based in part on the efficiency gains, as one officer commented 'I phoned up a witness in relation to a job, phoned them up and that job was in 2000, and she was amazed. I had all the dates, all the names and what have you at the touch of a button' but also on the actual image of police officers with technology; the same officer continued 'they are quite surprised at the level of our technology and they are pleased to see we are using that. That's the feedback I get from them [public].' A comment from a recent user was that the use of the MICT 'looks good to the public – but can take away a part of the social dimension of the role in terms of helping people with crime prevention issues and being prepared to listen. Not so easy to do that with your head stuck in a computer'. The creation of a positive image has been identified by the officers interviewed and is addressed in the literature reviewed, although the publics in the academic studies to date tend to be specific audiences rather than the police officers

'Joe Public'. These include the sales staff examined by Watad and DiSanzo (2000) and Okoli's work looking at academic conferences.

4.4 Equipment

This category is concerned with the ability of the equipment issued to meet the needs of the officers using it in the situations in which they use it. The category attracted relatively low numbers of comments but these were grouped into relatively few areas. This was the only category where the concerns about weaknesses or potential weaknesses strongly outweighed the expected strengths or potential strengths. Nearly three quarters of the comments made in this area were concerned with weaknesses or potential weaknesses and only one area attracted a positive set of comments – this was, however, concerned with the future abilities of the equipment rather than current.

The major area of concern for many officers centred on the issues of fragility and loss with many comments identifying fears of damage loss or theft of equipment. This was felt by one user to be an evolutionary issue in that 'ruggedized' versions of equipment would be developed and this user noted clearly that the technology per se was fine, it is the environment which is specific and renders it too fragile 'when you look at the successful stuff, kit that worked in that [police] environment, stuff from Husky, Psion which has been ruggedized – the degree to which it has been made operator proof and environment proof is quite high'. Some of this concern was based on the cost of the equipment but most on the potential security breach if confidential information was stored on the device, although the extent to which that could be protected was realized by many officers, as one commented 'I think the protection they have on the machines that they [informant handlers] have is too good for me to worry about that [theft of the equipment and loss of information]'. It is also useful to note that a minority concern was that damage or loss could be the result of user abuse of the equipment rather than genuine theft, or accidental or loss or damage. These concerns are reflected in other studies in public service environments such as education (Valiquette, 2000; Jipping et al., 2001; Soloway et al., 1999) as well as noting the limitations on input and output from the devices as identified by numerous writers in the field (Nielsen, 2002; Adam et al., 1997; Marcus, 2002). The issues of signal coverage and connections (Varshney 1999; Satyanaryanan et al., 1996; Ebling et al., 2002) are a major concern for users in this study with current users reporting significant problems both with via dial in and via GSM cards, partly as a function of signal strength and coverage. Overall officers felt that the connection technology was not as advanced as the hardware, and was not user-friendly.

The fear was expressed that the use of MICT would result in yet more cables and wires and it was also noted that the input for handheld devices is poor ('I'm not sure I could do that [stylus use] in gloves on a dark night in February') and that even laptop devices have keyboards which are less user-friendly than a full-size desktop device. The fear was expressed that handheld technology particularly could become a gadget or status symbol issued as a result of rank or functional specialisation rather than as a tool provided to meet a genuine job need and one officer noted the predilection of the police for such items 'if you take something small, black and shiny and you give it a three letter acronym then police officers will buy it' .The literature has charted the weaknesses and potential weaknesses of the equipment used in MICT quite fully (Satyaryanan, 1996, Varshney, 1999) and officers reflected many of these concerns. Portability was addressed by all users and there was a clear division between the perception of handheld devices and laptop equipment. The laptops were perceived as 'luggable' rather than portable and many comments identified the size and weight of the laptops as being a barrier to their routine use as mobile equipment. Almost all of the comments on PDAs were positive with relation to portability with the 'pocketable' nature of the devices being stressed – although it was also recognised that they are not just pocketable by the officers concerned – raising issues of loss of data and data security more generally. It should also be noted, however, that smaller is not always better – a point made by Krogstie (2003) in an analysis of the myths surrounding MICT use where he points out that there is a drive to smaller equipment where sometimes larger devices are more appropriate.

4.5 Infrastructure

This category is concerned with the support structures which underpin the use of the technology.

A high proportion the officers interviewed or observed expressed a concern that they have either not had, or are unlikely to get, adequate training to allow them to make the most the of the potentials of the technology with which they have been supplied. Officers felt that they needed variable levels of training with some identifying that they have good existing levels of facility with PCs and with equipment generally. One user commented that 'everyone is independent and individual and they will all use different things. It is a shame that some people don't use some of the things because of they knew how to do them it would make their jobs easier and it is only when you start to use them that you start to think "Yes, this is good" '. Officers also identified that they needed ongoing support and that a mobile user required qualitatively different support from that required by a desktop user – often needing the answer to a query instantly to allow an in-

progress task to be progressed. Communities of practice are seen as a solution to this and there are also informal solutions which have built up with current users where an individual or individuals are seen as 'super-users' ('I'm the IT helpdesk in here'). Other solutions include getting advice and support from friends and family as well as, in one example, from the victim of a crime. This issue of training is also supported in the literature and was a particular issue in the Watad and DiSanzo (2000) study. Watad and DiSanzo (2000) note that the implementation he described required 'several days of general PC training and specific software training' to start with and that after this formal training users started to carry out their own informal training. This training was sourced not just from inside the company but also outside from people such as friends and relatives. Watad and DiSanzo (2000) note that this may be an issue for security, as is also identified by Badamas (1999) and Lee and Lee (2002) who comment on the inherent risks of mobility and the role of user habits. A failure to train will, according to Juhlin (2000) lead to people being 'intimidated by the technology' and this is sometimes best addressed by being able to get support from colleagues.

5. AN INTEGRATIVE MODEL

This study has sought to examine the changes in work and working practices for police officers which have been produced as a result of the process of introduction of MICT into their working environment and routines. The study was based on a single force and did not seek to address the manner in which these changes occurred but, in the process of speaking with managers and with officers and of observing practice in this and in other forces we have started to develop an initial understanding of the process that is taking place and have included a statement of this below. This is an area where further work is needed and this paper is not in a position to address this although it is an area that the authors are keen to develop further.

In most police forces there is an initial planning stage for the introduction of mobile data – this may be a result of either top down pressure by senior managers or ICT departments or it may be a result of pressure from officers who have either seen technology deployed in other forces or have experience of the technology in other contexts and want to deploy a similar technology. Such pressure will lead the organization into some form of planning period during which they are likely to make an initial identification of business gain at a broad level, scope the technology and cost issues and then use this information to build a business case for a proof of concept or pilot. Such business cases vary dramatically from being argued and reasoned cases

based on force strategy and incorporating mapping of existing and proposed process to identify cost time and other gains through to being rough delivery plans aimed at scoping the issues involved and more at an experimentation level. From this stage forces will tend to move into a stage in which they concentrate on getting the chosen equipment and infrastructure to work – this may be done by an ICT department alone or by an ICT department working with a single supplier or a supplier consortium. At this stage forces will have identified some desired outcomes from the project – and these are usually in the form of incremental changes – improvements to existing ways of working, savings in travel or officer time and reduction in load on existing systems. This process is illustrated in Figure 1 below.

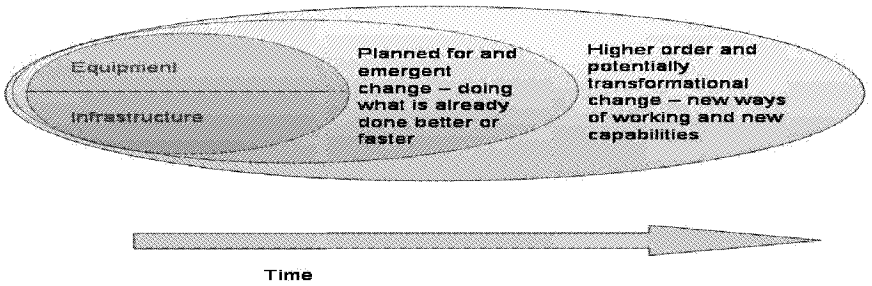


Figure 1. Stages of development and implementation of MICT in police forces

Figure 2 shows the process against a time line. This is not intended to provide a timescale but to indicate a sequence of events. The figure takes the process described above on a stage from the initial explanation above in that, as officers and other stakeholders implement their project and seek to manage the gains they have targeted and avoid the risks they have identified they may, and usually will, start to see ways in which this, or allied technologies can be used to transform the manner in which they work. This ties in with the original model proposed above to explain where effects can be seen in the work of officers and this is replicated in a similar format to figure 1 above.

As stated above, this is an initial and tentative explanation of the stages that an MICT implementation can go through in a police setting and is not fully developed. It is provided to offer some background and perspective to the main findings of the study as discussed in the paper, and to provide a basis for further discussion and research.

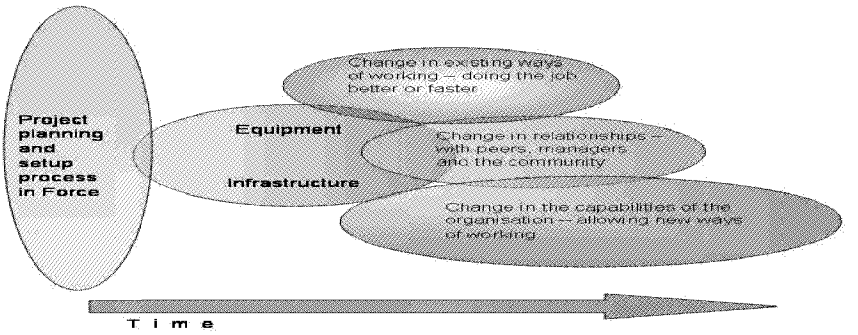


Figure 2. Areas of influence of MICT on police practice set against a broad timeline

6. CONCLUSIONS AND AREA FOR FUTURE RESEARCH

The conclusions for research are discussed below but it may be useful at this stage of the paper to signal the key areas that we feel to be of note for practitioners. The current study supports the idea that the police context can benefit from the use of MICT as others can and the study has highlighted some significant commonalities between the expectations that the police have of MICT and the effects experienced to date by the officers involved in their pilots of the technology. The police context is a major one in the UK and worldwide and the delivery of a public service of this type is a priority for this Government and others worldwide (Adderley and Musgrove, 2001). It also has parallels with other emergency services and contexts in the public and private sectors. In examining conclusions for practice there are two main areas of note and impact. The key point is that while the process of introduction of MICT into the police environment is one which has the potential to produce benefits for the officers, forces and communities involved these benefits will not be gained as a direct result of simply introducing MICT but will come from a process of introduction which will require planning and organisational adjustment as well as training and change for individual officers and teams. This is neither a simple process nor one which has any formulaic 'answer' which can be imposed. In examining the nature of this process the work done to date suggests that there are two 'hygiene factors' which need to be addressed before a

technology has any chance of being used – these are the areas of equipment and infrastructure (in the broader sense). If either of these have significant problems (poor connectivity, screens too small, input methods unsuitable, lack of training, poor technical support) then the technologies stand little chance of being used and moving into the mainstream of practice. Once these hygiene factors have been addressed there are three broad areas where the effects and impacts of the introduction of the technologies can be seen – in the relationships which users have with others, in the way that they carry out their work and in the addition of new capabilities to the way that the organisation works and the way that individual users work.

The study has highlighted a number of areas where further research may prove fruitful - ethnographic studies would help to provide fuller data on the nature of use and on the issues surrounding micro-mobility (Luff and Heath, 1998) and further real-life studies of this nature would allow the issues identified here to be extended and further explored. The matching of users to equipment and facilities is currently a quite broad process and it would be useful for further studies to examine the nature of the match between users and equipment or facilities. Luff and Heath (1998) have looked in detail at contexts and identified that there is a role for the evaluation of transactions at the level that they characterise as 'micro-mobility'. Such evaluations could lead to an examination of the factors which affect the use of the technology by individual users and the implications that this has both for the equipment provided and the manner of management of the change process generally.

Perhaps the largest single area for future work may be in the area of user involvement and the ISD methodologies which facilitate and incorporate this. The role of user involvement in the design of systems could usefully be clarified. Nandakhumar & Jones (2002) state that the 'literature on information systems almost unanimously recommends that users should be involved in the process of (IS) development', yet this does not appear to be a major feature of the design of systems using MICT according to the literature to date. This is, perhaps, part of a broader area for investigation which is the potential of information systems development methodologies to explicitly support the development of mobile information systems. The use of such systems is widely assumed to improve the quality of the systems produced; Walters, Broady and Hartley (1994) for example, comment that the use of a methodology 'helps to facilitate...the effective and efficient management of such (information) systems' but the literature identifies that relatively few organisation have such a methodology in place, and that even fewer report using it consistently; Fitzgerald (2000) notes that while many companies pay lip service to the idea of a methodology directing their development efforts close to 60% do not actually use one in the development they do. This study has shown that the police context in particular is a complex one and it may be that a broad approach such as Multiview (Avison

and Wood-Harper, 1990) or Soft Systems Methodologies (Checkland and Scholes, 1990) could support the development of more effective systems for the provision of information via MICT. User involvement is often promoted as a panacea for the development of systems which suit users and this is also suggested for MICT. However, as is always the case with user involvement this is a process which requires some care and precision and Fussell and Benimoff (1995) comment, '[we] have to involve users in the design of applications but also have to be aware that just asking' is unlikely to get all of the necessary detail''. There are also arguments that we need to get the views of non-users on the design of systems (Lindroth et al., 2000) in recognition of the fact that the use of MICT does have an impact on those around the users (Ling, 1999; Brodie and Perry, 2002).

References

Abowd, G. and Mynatt, E., (2000), Charting past, present and future research in ubiquitous computing, *ACM Transactions in Computer Human Interaction*, vol. 7, no. 1, pp. 29-58.

Adam, N., Awerbuch, B., Slonim, J., Wegner, P., and Yesha, Y., (1997), Globalising business, education, culture through the Internet, *Communications of the ACM*, vol. 40, no. 2, p. 115.

Avison, D., and Wood-Harper, A., (1990) *Multiview: An Exploration in Information Systems Development*, Blackwell Scientific Publications, Oxford, 1990

Badamas, M., (2001), Mobile computer systems - security considerations, *Information Management and Computer Security*, vol. 9, no. 3, pp. 134-136.

Bellotti, V., and Bly, S., (1996) Walking away from the desktop computer: distributed collaboration and mobility in a product design team, ACM . In Ackerman (Ed.) *Proceedings of CSCW '96* p.209-218, Boston MA

Brodie, J. and M. Perry (2001): *Designing for Mobility, Collaboration and Information use by Blue-Collar Workers*. Position paper presented at the Workshop on WORK/PLACE: Mobile Technologies and the Emergence of the New Workplace held in conjunction with the ECSCW 2001, Bonn, Germany.

Checkland, P., and Scholes, J., (1990) Soft systems methodology in action. John Wiley , New York, NY

Ebling, M., John, B., and Satyanaryanan, M., (2002) , The importance of translucence in mobile computing systems, *ACM Transactions on CHI*, vol. 9, no. 1, pp. 42-67.

Edvardsson, S and Bergqvist, J., (2000) Channel vs. Person-Orientation on Mobile Communication Devices. In Bertelsen, Bødker & Kuutti (Editors), *Proceedings of NordiCHI2000*, Stockholm, Sweden

Fails, J. and Olsen, D. (2002) Light widgets - interacting in everyday spaces, ACM, In Storey, Noy, Musen, Best, Fergerson and Ernst (Eds.) *Proceedings of IUI '02.*

Fitzmaurice, G. (2000), Situated information spaces and spatially aware palmtop computers, *Communications of the ACM*, vol. 36, no. 7.

Fussell, S. and Benimoff, N., (1995), Social and cognitive processes in interpersonal communication: implications for advanced telecommunications technologies, *Human Factors*, vol. 37, no. 2, p. 228.

Gadamer H., (1976) *Philosophical Hermeneutics*, University of California Press, California

Gaines, B.R. and Shaw, M.L.G. (1994). Concept maps indexing multimedia knowledge bases. *AAAI-94 Workshop: Indexing and Reuse in Multimedia Systems.* pp.36-45. Menlo Park, California, AAAI.

Green, N. (2002) On the move: technology, mobility, and the mediation of social time and space. *Information Society* 18(4) pp.281-92

Guerlain, S., Lee, J., Kopischke, T., Roamnko, T., Reutiman, P., and Nelson, S. ,(1999), Supporting collaborative field operations with personal information processing systems, *Mobile Networks and Applications*, vol. 4, pp. 37-48.

Hauck, R. and Chen, H. (1999) COPLINK - a case of intelligent analysis and knowledge management. In the *Proceedings of International Conference on Information Systems* (ICIS '99) Charlotte, NC, Dec 1999

Howcroft, D and Hughes, J., (1999) Grounded theory – I mentioned it once but I think I got away with it. In Brooks, L and Kimble, C (Eds.) *.Proceedings of the 4th UKAIS Conference* Basingstoke, McGraw-Hill, York 1999, pp 129 -142

Hughes, J., and Howcroft, D., (2000) Grounded Theory: Never Knowingly Understood *Information Systems Review.* Vol 1, 181-197

Jipping, M., Krikke, J., Dieter, S., and Sandro, S., (2001), Using handheld computers in the classroom: laboratories and collaboration on handheld machines, ACM, In Henry MacKay Walker, Renée A. McCauley, Judith L. Gersting, Ingrid Russell (Eds.): *Proceedings of the 32rd SIGCSE Technical Symposium on Computer Science Education, 2001*, Charlotte, North Carolina, USA, 2001. ACM 2001 .

Juhlin, O. and Normark, D., (2000) Bus driver talk. In Svensson, L., Snis, U., Sørensen, C., Fägerlind, H., Lindrotha, T., Magnusson, M., and Östlund, C., (editors), *Proceedings of IRIS 23.* Laboratorium for Interaction Technology, University of Trollhättan Uddevalla, 2000

Kakihara, M. and Sørensen, C., (2002a), Post-Modern' Professionals Work and Mobile Technology. In *Information Systems Research Seminar in Scandinavia (IRIS'25)*, Denmark. Copenhagen Business School.

Kakihara, M. and Sørensen, C., (2002b), *Mobility: An Extended Perspective.* In Thirty-Fifth Hawaii International Conference on System Sciences (HICSS-35), Big Island Hawaii

Kakihara, M., C. Sørensen and M. Wiberg (2002). Fluid Interaction in Mobile Work Practices. In *Proceedings of the 1st Tokyo Mobile Roundtable, Mobile Innovation Research Program, Institute of Innovation Research*

Kakihara, M. and Sorensen, C., (2001) Mobility Reconsidered: Topological Aspects of Interaction , *Proceedings of IRIS24.* Ulvik, Norway: Vol. II, pp. 99-114.

Kristoffersen, S., and Ljungberg, F., (1999c), Making place to make IT work: empirical explorations of HCI for mobile CSCW. ACM. In *Workshop on Groupware related Task Design at GROUP'99 Conference, Phoenix, Arizona, USA* , November 1999.

Kristoffersen, S. and Ljungberg, F., (1996) *Darwin and Newton - basic infrastructural information technologies for group work.* In the BIIT project. edited by Beck, E., Department of Informatics, University of Oslo.

Krogstie, J (2003) *Mobile process support systems:myths and misconceptions* . Extended abstract for the for workshop on Ubiquitous Computing environment, Weatherhead School of Management, Case Western Reserve University, Cleveland, Ohio, USA K Lyytinen and Y Yoo (eds) at http://weatherhead.cwru.edu/pervasive. Online 02 Feb 05

Laurier, E., and Philo, C., (1998). *Meet you at Junction 17: a socio technical and spatial study of the mobile office.* Glasgow, Dept. of Geography, University of Glasgow and ESRC, Swindon.

Lindgren, R. & Wiberg, M. (2000) Knowledge Management and Mobility in a Semi-Virtual Organization: Lessons Learned from the Case of Telia Nära. In Sprague,R (Ed.) *Proceedings of HICSS-33*, IEEE-press. Maui, Hawaii, January 4-7, 2000

Lindroth, T., Nilsson, S., and Rasmussen, P.(2001) Mobile usability: rigour meets relevance when usability goes mobile. In Bjørnestad, S. et al. (eds.).*Proceedings of IRIS24*, Ulvik, Norway

Ling, R (1999) Restaurants, mobile telephones and bad manners: New technology and the shifting of social boundaries. Paper presented at *Human Factors in Telecommunication 1999*, Copenhagen, Denmark.

Luff, P. and Heath, C., (1998) Mobility in Collaboration. in Poltrock, S., et al. eds. Proceedings of *CSCW 1998*, ACM Press, Seattle, Washington, 1998, 305-314.

Marcus, A. and Chen, E., (2002), Designing the pda of the future, *Interactions*, vol. 2002, Jan-Feb, p. 34.

Nandakhumar, J. & Jones, M. (2002) Designing in the dark - the changing user developer relationship in information systems development. 2002. Proceedings of the *18th International Conference on Information Systems*. ACM

Nielsen, C. and Sondergaard, A., (2000)Designing for mobility - providing integration and overview on large and small screens, In Svensson, L., Snis, U., Sørensen, C., Fägerlind, H., Lindrotha, T., Magnusson, M., and Östlund, C., (editors), *Proceedings of IRIS 23*. Laboratorium for Interaction Technology, University of Trollhättan Uddevalla, 2000

Nishibe, Y. and Waki, H., (1998), Mobile digital assistants for community support, *AI Magazine*, vol. 19, no. 2, p. 31.

Nulden, U. (2003) Police Patrol Mobility. Abstract of Paper presented at the *Case Western Workshop on Ubiquitous Computing 24-26 October 2003*. Online, 01 Feb 05 at http://weatherhead.cwru.edu/pervasive/participants_one.htm

Okoli, C., Jessup, L., Ives, B., and Valacich, J., (2002) Mobile conference information system: unleashing academic conferences with mobile computing, IEEE, In *Proceedings ICIS 02*.

Orlikowski, W., (1993) *CASE tools as organisational change: investigating incremental and radical changes in system development.* MIS Quarterly 17(3) pages 309-340

Pakanen, J., Mottonen, V., Hyytinen, M., Ruonansuu, H., and Tormakangas, K. (2001) A web based information system for diagnosing, servicing and operating heating systems. ITCon 6, 45-55 Consulted online at http://www.itcon.org/2001/4 August 2002

Pica, D., C. Sorensen, & D. Allen (2004): On Mobility and Context of Work: Exploring Mobile Police Work. In *Thirty-Seventh Hawaii International Conference on System Sciences (HICSS-37),* Big Island Hawaii, ed. R. Sprague Jr. IEEE. www.hicss.org

Pica, D. & M. Kakihara (2003): The Duality of Mobility: Understanding Fluid Organizations and Stable Interaction. In Mercurio,R (Ed.) *Proceedings of ECIS 2003*, Naples, Italy.

Satyanaryanan, M. (1996) Fundamental challenges in mobile computing, ACM, In Jansen, W (Ed.) *Proceedings of PODC 96.*

Soloway, E., Grant, W., Tinker, R., Roschelle, J., Mills, M., Resnick, M., Berg, R., and Eisenberg, M., (1999), Science in the palms of their hands, *Communications of the ACM*, vol. 42, no. 8, p. 21.

Sorensen, C. (2003) Research Issues in Mobile Informatics : Classical concerns, pragmatic issues and emerging discourses. Position paper for workshop on Ubiquitous Computing environment, Weatherhead School of Management, Case Western Reserve University, Cleveland, Ohio, USA K Lyytinen and Y Yoo (eds) at http://weatherhead.cwru.edu/pervasive. Online 02 Feb 05

Vaast, E. and Walsham, G (2005) Representations and actions: the transformation of work practices with IT use. *Information and Organization* 15(2005) 65-89

Valiquette, P., Miller, M., and Seeger, E., (2000), Mobile computing at Renesselaer Polytechnic Institute, ACM, In Reimer, U (Ed.) *Proceedings of SIGUCCS 00.*

Varshney, U., (1999), Networking support for mobile computing, *Communications of the AIS*, vol. 1, no. 1.

Walters, S., Broady, J., and Hartley, R., (1994) A review of information systems development methodologies. *Library Management* 15 (6) 5-19

Watad, M. and DiSanzo, F., (2000), Case study: the synergism of telecommuting and office automation, *Sloan Management Review*, vol. 41, no. 2, p. 85.

Weilenmann, A., (2003) "I can't talk now, I'm in a fitting room": Availability and Location in Mobile Phone Conversations, in *Environment and Planning A* volume 35, (9) September, pages 1589 - 1605, special issue on Technology and Mobility, ed. E. Laurier

Weilenmann, A. (2001) Mobile Methodologies: Experiences from Studies of Mobile Technologies-in-Use, Published in Proceedings of the 24th Information Systems Research Seminar in Scandinavia (IRIS 24), Bjørnestad et al.. (eds) vol. 3: 243-257

Wiberg, M. and Gronlund, A., (2000), Exploring mobile CSCW: 5 areas of questions for further research . In the proceedings (Svensson, L., Snis, U., Sørensen, C., Fägerlind, H., Lindroth, T., Magnusson, M., & Östlund, C. (Eds.)) of IRIS 23, August 12-15, 2001, Uddevalla, Sweden

Yoo, Y. and Lyytinen, K. (2003) Measuring the consequences of ubiquitous computing in a networked organization. Abstract of paper for the workshop on Ubiquitous Computing environment, Weatherhead School of Management, Case Western Reserve University, Cleveland, Ohio, USA K Lyytinen and Y Yoo (eds) at http://weatherhead.cwru.edu/pervasive.Online 02 Feb 05

THE KNOWLEDGE AND THE SYSTEM

A socio technical view for supporting London Black Cab Work

Silvia Elaluf-Calderwood and Carsten Sorensen
London School of Economics and Political Science; Houghton Street, London WC2A 2AE, UK
{s.m.elaluf-calderwood, c.sorensen}@lse.ac.uk

Abstract: This paper reflects on the socio technical implications of two different technology-based Black Cab booking systems. Potentially there is a bi-directional impact on the drivers and passengers with respect to the level of awareness needed to use the systems and how situational acts vs. planned acts impact on the context changes experienced by all users.

Keywords: Situational Acts, Planned Acts, Mobile ICT, booking systems, cab

1. INTRODUCTION

The use of mobile technologies supporting work is currently the subject of significant research in many areas of work such as police work and health services (Sorensen and Pica, 2005; Wiredu, 2005). One of the key areas of concern is the relationship between emerging working practices and the specific properties of mobile ICT support. The mobile nature of the technology clearly strengthens the ties between the work situation, the possible decisions made by the worker, and the specific design of the technological support. Furthermore, it potentially reconfigures the ties between the individual worker and the organisational context in which work is conducted.

One interesting aspect of studying London Black Cabs in the light of mobile technologies supporting work is the possibility of comparing traditional working practices settled for many years. London Black Cabs are heavily regulated and their history goes back to 1620 when they were called Hackney cabs- such as street hailing (Bobbit, 2002) against modern work

practices such as allocation of work by SMS or electronic booking systems. A second interesting aspect of the study of London Black Cabs is identifying the importance for drivers of relying on "The Knowledge" (a memory system for street routes that drivers are examined on by the Public Carriage Office in order to obtain their license) against relying on GPS systems or other "live" modes of data stream for location services. The use of location services or GPS is an ongoing process that it is producing dramatic changes, by means of, mobile ICT in the everyday working practices for drivers.

An important question is then: *What is the role of specific mobile ICT support features for the change to daily working practices of London Black Cab drivers?* This question relates to a more broad research concern: understanding the role of mobile ICT for the choice of work context for mobile workers. The decision process that arises from the choice between planned and/or situational acts.

The empirical data for this paper is provided by both qualitative interviews with 35 black cab drivers from whom deep contextual knowledge was gained, and through 14 hours of videotaped observation of driver-behavior (this part of the research is still being developed) in order to obtain deep situational insight. This paper in particular focuses on a comparison between two different technological means for connecting a potential customer to a black cab within the normal working conditions of the driver.

Although still in its early stages, there has been some research of the socio-technical aspects of mobile working (Orr, 1996; Kristoffersen and Ljungberg, 2000; Brown et al., 2001; Wiberg, 2001; Kakihara, 2003; Ling, 2004; Wiredu, 2005). In particular studies of police work in patrol cars (Manning, 2003; Sørensen and Pica, 2005) and of airplane pilots (Hutchins, 1995) may be relevant for the understanding of vehicle-based working.

The paper describes some of the specific properties for the two mobile ICTs studied in terms of the systems allocation of work to cab drivers, design factors in the system and the discourses related to the mobile ICT practices of emerging versus planned decision processes.

2. THE BLACK CAB, THE KNOWLEDGE AND THE SYSTEM

2.1 The London Black Cab Service

The history of the London Black Cab is rich and long (Georgano, 2000 and Bobbit, 2002). Three main reasons lead us to identify this type or work as an interesting case for study: it is still possible to compare live data between established ways of work and new technology driven ways to do work, the use of "The Knowledge" and changes due to technology modifying the way work was traditionally executed by cab drivers to use the mobile ICT. In the context of the London Metropolitan area, the use of electronic booking systems in cabs is a relatively new development. Unlike many other cities in the world where the use of GPS systems is generalised in cabs (Liao, 2003), most radio cab circuits still use inherited systems based on two-way radios and paper-clip bookings. This is the case for both licensed cabs and minicabs (restricted licensing).

In terms of operation modes there are two main types of cabs: licensed cabs and minicabs. The differences are based in the licensing method, the fares, and requirements for route planning. London Black Cab drivers, to become fully licensed, need to pass "The Knowledge", an exam that requires them to know and recall from memory up to 400 routes or "runs" in Greater London. Black Cab drivers or "cabbies" (Townsend, 2003) are proud of this standard of knowledge that allows them so far to be faster than any GPS systems available (Skok, 1999).

Most cabbies own their vehicles and are proud of their high level of independence when choosing work. Minicab companies, however, have been able to compete with cabbies by hiring drivers who do not have "The Knowledge" but can complement their routes using GPS systems. Minicab drivers tend to drive vehicles owned by the minicab company and they are much less independent as the cabbies when choosing work.

The migration to electronic cab booking systems aims to take maximum advantage of the position of the cab at a certain time. Most radio circuits see this position awareness as a strategic advance when allocating work. When a booking is made, the job is allocated to a driver close to the passenger, reducing the arrival time (of the cab to the passenger) and waiting time (by the cab driver). Cabbies do compete with each other for hailing passengers, hence members of the same radio circuit tend to be overzealous when determining whether the allocation of the job has been fair and not subject to the call centre dispatcher preferences. The use of computerized booking

systems is a means both for optimization and for reducing possible conflicts between drivers.

Based on the empirical data and in a socio technical approach here is a list of some factors to consider in the design of the electronic booking systems:

- *Ubiquity*: identifying the location of drivers closer to the passenger wishing to be transported. This can be an accurate position (by GPS) or an estimated position (by zones).
- *Reachability*: communication between cab driver and call centre or cab driver and passenger shall be intelligible enough to provide basic information about ride.
- *Security*: passengers and cab drivers value the ability to travel and drive in safety. An added value for any system is the capacity for providing a backup communication media in the case of emergencies.
- *Ergonomics:* from the driver's perspective, a system that minimizes distractions from driving concentration is important; it is also important to note that the billing systems (credit card swap or account register) can also be incorporated into the system.
- *Easy learning:* from the driver's point of view it is important that the system should not be difficult to use, should be easy to understand and allows the rapid location of relevant information. The complexity of this objective is increased by the fact that drivers have different levels of general education and computer knowledge.

The role of these factors contributes to the mechanism of doing mobile work, when merging its role with the technology in use. Those mechanisms for doing mobile work are explained in the section 2.2

2.2 The Mechanism of Doing Mobile Work

The cab driver work can be presented from the interaction of the physical space, the mobile actor and the technology attached to work (Weilenmman, 2004, Elaluf-Calderwood and Sørensen, 2004). This way of looking at cab drivers' work is complemented by understanding the idea of what mobile work is and what it means in the context of spatially mobile workers.

From the interviews and observation completed it can be said that when a cab driver works around the city searching for work, the search and its success[1] depends on a number of factors:

[1] During interviews drivers have expressed different aims when measuring their workday success. For example: some drivers aim to make as much money as possible depending in their personal situations, others want to do easy rides (short journeys), other want to have a relax driving day with many breaks, etc.

- *Physical location*: where the driver and his vehicle are physically located at a certain time.
- *Awareness*: The driver needs to be aware of events on the road such as accidents, congestion, competition from other drivers, etc.
- *Time*: drivers go to work with a general timetable framework; drivers perceive time either as a compressed unit (E.g.: driving whilst talking to friends) or fragmented (E.g.: events or actions can occur at discrete intervals of time (E.g. one conversation followed up between two passenger hails using the mobile phone).
- *Strategic Planning*: cab drivers' decisions such as how many jobs they wish to take from the electronic booking system, when they will take those and where. E.g. some drivers plan to only take hails in the direction of the drivers home one hour before finishing their work shift, allowing them to get close to home whilst being paid for it.
- *Situational Acts*: when on the road which hailing situations are preferred by drivers based on the context of work, but also the opposite (not preferred and why)
- *Planned Acts*: different from strategic planning as the time intervals when these acts are planned are short and are a function of the randomness of available work.
- *Human Factors*: How the cab driver mediates with the mechanical, technological and human aspects of his work. If the driver is tired, lonely, stressed, or subject to other human emotions, it will affect the way jobs are taken.
- *Role of the technology*: cab drivers might use one preferred system for obtaining jobs or might choose to be more traditional and work the street hailing of passengers.
- *Emergent practices*: the evolution of the physical space in the city together with new technologies is creating new working practices that cab drivers are taking on board for their work.
- *Change to succeed*: this category expresses the randomness of the work, as success might occur "at the first turn of an imaginary fortune wheel" or might take many of the factors listed above in combination. Drivers express their measure of success in different ways currently under analysis.

Based on this socio technological approach a brief introduction to the research approach is presented in the section below before passing to discuss the Systems observed based on the factors in the mechanism of doing mobile work.

3. RESEARCH APPROACH

The research approach is based on interpretivism and ethnographic methods. The interpretative approach tries to understand the world as it is, created by inter-subjective meanings in a social process. It tries to understand a social phenomenon from the perspective of participants in their natural setting. In an interpretative study, the researcher does not try to impose his/her own previous understanding onto the situation. The case study (London Black Cab) presented in this article places a significant interpretative value on the narrative as expressed by the object of study (the subject is being interviewed and what is said in the context of study). The understanding of the differences between the living situation of the mobile user and the researcher's hypothetical views is of paramount importance. Distinguishing between situated interaction in the world on one hand and interaction thought technologies on the other is at the heart of virtual environments and mobility studies (Luff and Heath, 1998). The strategy defined for the collection of data was the use of one-one interviews with drivers from diverse social and cultural backgrounds. There has been research work (still ongoing) in recording everyday situations in which drivers use their computer systems and/or mobile devices. For analytical techniques the research is being worked using cognitive approaches, and microanalysis has been applied when required. There is very limited access to logs or records provided for the systems discussed, hence *this paper focuses on how the systems are used* (cab drivers, passengers) and *not on how the systems work* (interfaces, mobile antennas, software, hardware, etc).

4. TWO WAYS OF ALLOCATING JOBS

4.1 Driver-Passenger Interaction

We now aim to describe the interaction between driver and passenger in System A and System B.

System A

This is a cab radio call system that has fully migrated to an electronic booking system. Potential passengers book by phone or over the Internet (no SMS accepted). Payment methods are of three types: cash, credit card and account. The company gives preference to account customers over all other

transactions. Customers get an ID number for their booking or an email confirmation. An estimated cost of the journey is also provided on request and customers are required to state their final destination for this purpose. Customers do not interact with the cab driver until they are picked up at the start of the journey.

Drivers know the location of the customer and their approximate destination. Drivers book in the electronic system based on a virtual zone map. They manually input the zone they wish to be included in. The system automatically assigns them to a virtual queue – based on order of input of zone – of cabs. To avoid driver "choose and pick" of jobs from the system, drivers cannot see the destination (or the cost) until the job is accepted. Drivers can have three modes of "location" in the system (if on): p.o.b (passenger on board), free (no passenger), c.t.e (close to end of journey, which allows the driver to be re-assigned a place in the queue). There is also the option to be off the system if hailing passengers from the street.

Discrepancies do occur, as passengers are charged from the time the cab arrives to the point of collection, regardless of how long the passenger makes the cab wait, so for the last three years the new computer systems are being provided with a GPS system that can be checked by the call centre when there is disrepute. In general the service is good, drivers tend to work a combination of radio jobs and street jobs.

Drivers have mobile phones in their cabs but it is not the primary mode of communication for work. Most work allocation is negotiated by the computerized system.

Drivers feel very comfortable with their radio circuit passengers as in most cases these will be account passengers, what they call "quality people" (business people), hence there is a level of relaxation associated with having knowledge of the journey destination.

Drivers express a level of respect for their passengers in their behavior: trying to make few or no calls on their mobile phones or keeping the radio volume low. Drivers are very receptive to passengers' attitudes and desires (if the passenger wants to chat, the driver will listen and try to be more communicative). This is also true if dealing with cash jobs at late nights, as far as the booking is made through the radio circuit.

System B

The cab is provided with a GPS system that sends the location of the driver in real time to an application server. Passengers call a number using their mobile phone and based on the location services of the network the closest cab to the passenger is called directly through the mobile phone in the cab. The passenger is, the closest cab to him/her is called directly

(language) and cab driver and passenger negotiate by voice call where the passenger will be picked up, the estimated time of arrival at the destination (for the cab) and in some cases where the passenger wants to go and method of payment.

Street pick ups or street hails can also be booked by SMS, in which case the driver calls the passenger to confirm call – there is no Internet or call centre available (at this moment in time) – using the text header in the SMS message.

The main issue for drivers is that customers do not wait for the cab they booked if there is another available arriving earlier on. This makes the drivers wary of customers disappearing. It is perceived as a disadvantage that the customer cannot be charged a deposit in advance for the assignment. There are also fluctuations in the precision of the GPS systems and how jobs are allocated – there is no queuing – hence the driver has the added tasks of getting information from the passenger, defining the driving route, driving with the passenger and obtaining his/her payment.

Drivers use their mobile phones as the primary source of communication for job allocation. Job rejection does not affect the chances of getting a new job allocated immediately. However if the driver answers the call he is obliged to comply and do the run (travel to the passenger).

This issue depends upon the time of the day. At nights drivers are at especially high risk of not finding passengers or not being paid. Hence drivers have a more reserved attitude to jobs and tend to make rounds around the city more frequently than the drivers in System A.

There is no call centre associated with the system, hence if a driver gets in trouble the only assistance he/she can reach is using his/her mobile phone. The GPS system is used to locate drivers but does not help them to find the optimum route to their destination. This is primarily planned as in System A using the knowledge - the very difficult examination taken by London cab drivers in order to get their license with the Public Carriage Office. In figures 1 and 2 below we present physical pictures of the systems described.

Figure 1. System A

Figure 2. System B

4.2 Systems Comparison

We now describe the major characteristics of the systems and give examples of how operations are completed.

Table 1. Socio technical description of the systems

	System A	System B
Ubiquity	Driver has the system integrated in the vehicle. Driver determiners when the cab is made available. Driver is company shareholder with participation in the working practice decisions. Min 30 rides a month are required.	Driver has the system integrated in the vehicle. The driver determines when the cab is made available. The driver pays a flat free (monthly) for access to the system. No min number of rides.
Reachability	Multiple repeaters around London. Good reception. The system has a manual backup and alternative backbone network.	Access is supported by commercial satellites and GPS network. Good receptions but sometimes there are some reachability issues.
Security (from competition)	Good. Access to transactions and locations for driver and passenger are managed by call center. Once a booking is completed the chances that the job will be taken away from allocated driver is low, Drivers only can see non-allocated jobs. In most cases passengers need to provide a credit card for payment.	Reasonable. Communications between driver and passengers are completed using digital mobile networks. There is no assurance that once driver and passenger have agreed to the service, the passenger will wait for cab. Sometimes passengers do take the first cab that arrives to their position and not the one booked.

		System A	System B
Convenience	Driver	Driver is given in advance all information for ride by computer system: passenger location, destination, cost and payment.	Driver has to negotiate with passenger collection time, location, are and method of payment. This negotiation can distract the driver from driving well when on the road.
	User	Passenger can book cab by phone call or internet. Methods of payment are diverse. Passenger is also provided with estimated time of arrival for booked cab.	Passengers can only book by mobile phone calls. External conditions such as noise of the road can affect the quality of the call - can take long to get passenger details - and in some cases another cab will arrive to passenger location and passenger will cut call canceling the booking.
Localisation		Driver inputs his location in the Zone system. GPS is not used to verify the position of the driver unless there is a complaint or dispute. The driver can change the zone system manually at any time.	Driver's location is determined by an advanced GPS system build into the vehicle. This system is a real time feature that cannot be changed manually but only turned off.
Instant Connectivity	Driver	Driver is able to correlate a position in a cab queue when waiting for a new job assignment.	There is no queuing system. The nearest cab to the passenger gets the job from the system
	User	A reply is obtained from a call centre or internet page and email confirming booking.	Direct communication with driver and confirmation that its on its way.

* From the driver point of view unless otherwise stated.

5. HOW DOES ICT SUPPORTS BLACK CAB WORK

5.1 Emerging versus Planned Decisions

When using System A, the passenger exchanges instant connectivity for convenience (if the cab booked does not arrive on time, the radio circuit is able to provide the next closest one in a matter of minutes). When using System B the passenger has a level of ubiquity attached to his position, and security is exchanged for instant connectivity (the passenger and the driver talk to each other in real time to discuss when and where to be collected).

This way of negotiating position is not unique to the cab business; the police also use diverse methods of communication to create awareness of position and location, and there are similar trade-offs between these methods and instant connectivity (Sørensen and Pica, 2005). In order to understand this without undermining the factors that attract passengers to choose one method over another, we also need to look into the driver's convenience and the ideas provided by mobility studies (Perry et al, 2001).

As the cab driver moves around the city searching for possible passengers, whilst being shown as available for the radio circuit, many events might occur. Radio circuit, the generic name used for computerized cab circuits, is seen as a reliable – but not the main or only - source of income for cab drivers. This is in part due to the careful control of the cost of each journey, competing against street hails, which are less carefully recorded and where discrepancies can occur.

With system A, in which the driver relies on the information provided by the computerised system to obtain work, each time he is "live" in the system he is allocated a queuing number. This queuing number allows the driver some level of planning (the question: which job I will take?) based on his vehicle physical approximate location within the parameters of the system (zones). In some cases for example, a cab driver might have a queuing number such as "4", and whilst waiting to ascend to the top of the queue, the driver might decide to take a short run around a physical area or stop at a cab rank for a break.

The cabbie is more relaxed as the pressures of constantly searching for new passengers is reduced by the greater trust placed in the computer system.

Cab drivers in system A know in advance their destination or proximity even before they have collected the passenger, allowing them to check routes, verify that there are no road closures, etc. There is a level of safety associated with the idea of traveling when the destination is known.

With System B, in which the driver relies on his mobile phone to obtain work – besides street hails – the ubiquity is wider. Drivers get accustomed to longer runs on specific routes to maximise the number of passengers transported. However passengers sometimes take the first cab that is closer to them and the driver looses his ride.

During the interviews drivers in System A, those were quoted saying: this uncertainty is the main reason they felt discouraged from trying the system B.

With System B drivers argued that using the system was advantageous when working at night: the cab density (number of cabs available) is reduced, passengers are more eager to confirm that the cab is a licensed one (especially female passengers) and are prepared to wait longer times – if

necessary -until in arrives. There is hefty competition from minicabs but black cabs feel backed by their good reputation.

There is a relationship between the spatial distribution of cabs and the passengers that affects the social interaction, the expectations (from drives and passengers) before the journey, during the journey and at the end of the journey.

With System B drivers find more difficult to control their petrol costs: they tend to aim to find the information required for collection of the passenger, but only on collection find out the final destination. Sometimes the journey to the passenger can take as much time as the hail itself due to traffic congestion. Uncertainty becomes part of the space of interaction between the driver and passenger. At some point the private space of the cab driver is claimed as public by the passenger (during the journey); at the same time there is an interaction that links the situational acts (collecting a passenger) with the planned ones (where to go, cab driver choosing to be available or not), etc.

5.2 Systems and their Riddles

Cab drivers concerns are driven by a sense of risk attached to the idea of multiple tasking when driving their vehicles, being wary of multitasking when working. Hence the need of a simple electronic booking system.

In system A, drivers claimed that the computer screen was a distraction to their driving, and liked the fact that the system turns off to black screen after two minutes idle. They perceived system A as fair, with litle competition between drivers in the radio circuit.

Nevertheless problems do occur, such as when drivers are on the boundary of one zone and the job allocated is too far within the zone or there are physical obstacles that make the journey not worthwhile (sudden closure of roads or a one way system), and there is also the probability that another driver is closer to the passenger to be collected. Then communication to the call centre is required to clear doubts. Misunderstanding can also arise when the description for the collection point or the passenger to be collected is not clear.

In system B drivers are concerned with the accuracy of the GPS systems used. Central London has a high density of mobile masts; hence the accuracy of their most probable position is high. However, there is also greater competition: since the system does not provide a queuing system, two drivers in the same street might both be ideally located for a job or run appearing in the street, yet the allocation of the job is random and there is always a chance that the passenger will take the first taxi that passes close to him regardless of the agreement they might have with another driver.

There are also issues concerning billing and payment. Taxi drivers in system A know in advance the method of payment (cash, account, credit card) and since their destination is also known they have an advantage when estimating the best route and cost.

Failures occur in the computer systems when no jobs can be allocated. If the system is down notification is provided to drivers. For drivers in system A, a broadcast in the two-way radio system announces the problem. For drivers in system B, a broadcast SMS is the way of announcement.

This type of system failure does not affect the passenger's potential to get a cab when needed, but it might affect the payment method (if paying by credit card). It will also affect the interactivity between driver and passengers as uncertainty is added to the journey in terms of ubiquity and reachability. In these situations some drivers switch to street mode until they are sure that the systems do work properly again. Having said this, outages are infrequent.

Computers in vehicles may fail and in that case both systems have workshops where cabs can be repaired. This activity implies a down time off the street, which drivers find difficult accept as it is unplanned time of the road that has a cost in their day profits.

5.3 Is Mobile ICT Challenging Mobile Work ?

The first and main concern for drivers is the competition between drivers using the Knowledge and drivers using GPS systems, or in other words licensed cabs vs. unlicensed cabs (for now). GPS systems in London are not yet able to provide "real time" information of what is happening a the very congested London roads (Example: routes closed by the police due to an emergency are only updated in GPS systems after a gap in real time). Cabbies are aware that this will not be in the case in future when technology will be able to provide location services with "real time". The need for the Knowledge is then questioned not only because of the technology but also by other factors such as congestion, more routes defined as one-way systems, many alleys and shortcuts being closed due to safety measures.

Drivers express this change as a way of making their "skilled" job an unskilled one; anyone with a GPS could do their job. No specific training will be required to do a cab job. This is the case with minicab companies, which are gradually obtaining a bigger share of the cab market. Their costs are reduced because those companies can hire drivers at low rates, who are not required to own their cabs for work.

Cabbies see the use of mobile technology as an enhancement of their private and social life, which can continue even when at work. They also appreciate the fact that their passengers seem to be at ease in using the back

of their cabs as an extension of their offices, homes, bars, etc by using mobile technology to be in touch with whom they want. Cabbies do highlight that as their work is isolated, human interaction is achieved through their mobile phone.

6. CONCLUSIONS

This paper has not aimed to do a theoretical review of the design parameters used in the design of cab systems, neither to describe state of the art systems but to analyze how everyday cab drivers adapt their working practices depending upon the technology.

In this paper only part of the empirical work completed has been presented. There still a considerable amount of data to be analyzed under the socio technical lenses explained in the introduction to this document. Considering the complex spectrum of issues related to time and space faced by cab drivers and their passengers, there remains considerable research into how new technologies improve the services provided by the cab drivers to passengers. A first questions regarding the evolution of human-to-human resource knowledge (such as the Knowledge) against computer-to-human resources (GPS systems) used by cab drivers arises: Will "The Knowledge" be replaced by more advanced GPS systems? Will call centres become redundant if smart systems could automatically handle passenger bookings? How will this affect the passenger and driver expectations of the service provided and used?

There is from the social pint of view a richness and variety on the cab driver's job as per changed through the use of mobile ICT. Drivers expressed feelings of isolation when doing their everyday work. Mobile ICT use is allowing drivers to overcome their isolation.

In terms of understanding the factors that are relevant for mechanisms for executing work with mobile workers, this ongoing research expects to contribute to further research work with the development of a model that maps the factors listed in Section 2 with the occupational frameworks of time and space, currently being blurred through the use of technology such as mobile phones and doing so mobile workers try to make sense of the socio technical issues arisen by the use of mobile ICT.

References

Bobbit, M., 2002, *Taxi: The Story of the London Taxi Cab.* UK, Veloce Publishing PLC.

Brown, B., N. Green, & R. Harper, ed., 2001, *Wireless World* Springer-Verlag UK.

Elaluf-Calderwood, S and Sørensen, C, 2004, *Mobile Work-Mobile Life*, 5th World Wireless Conference Proceeding, University of Surrey, UK

Georgano, G. N., 2000, *The London taxi.* UK, Shire Publications Ltd.

Hutchins, E., 1995: *Cognition in the wild.* Cambridge, MA: MIT Press.

Kakihara, M. ,2003, *Emerging Work Practices of ICT-Enabled Mobile Professionals.* PhD Thesis. The London School of Economics and Political Science. http://is.lse.ac.uk/research/theses/

Kristoffersen, S. & Ljungberg, F., 2000, *Mobility: From stationary to mobile work.* In *Planet Internet*, ed. K. Braa, C. Sørensen, and B. Dahlbom. Lund, Sweden: Studentliteratur, pp. 41-64.

Liao, Z., 2003, *Real-Time taxi dispatching using Global Positioning Systems.* Communications of the ACM 46(5): 81-83.

Ling, R.,2002, *The social juxtaposition of mobile telephone conversations and public spaces.* International Conference of the Social Consequences of Mobiles Phones, Chuchon, Korea.

Ling, R., 2004, *The mobile connection : the cell phone's impact on society.* Amsterdam: Morgan Kaufmann.

Luff, P. and Heath, C., 1998, Mobility in Collaboration. Proceedings of ACM 1998 Conference on Computer Supported Cooperative Work, ACM Press.

Manning, P. K., 2003, *Policing contingencies.* Chicago: University of Chicago Press. 0226503518 (cloth alk. paper). Peter K. Manning.

Mason, M., 2003, *In a Year of a London Cabbie Everyone has a story.* London, UK, Orion Publications.

Orr, J. E., 1996, *Talking About Machines: An Ethnography of a Modern Job.* Cornell University Press.

Perry, M., K. O'Hara, et al., 2001, *Dealing with mobility: understanding access anytime, anywhere.* ACM Transaction on Computer-Human Interaction (TOCHI) 8(4): 323-347.

Pica, D. and Sørensen, C. 2003. *On Mobility and Context to work: Exploring mobile Police Work.* London, London School of Economics.

Schmidt, K., 1998, Some notes on mutual awareness. COTCOS Awareness SIG Workshop, Paris, France.

Schmidt, K. and C. Simone, C., 1996, *Coordination Mechanisms: Towards a Conceptual Foundation of CSCW Systems Design.* Computer Supported Cooperative Work: The journal of Collaborative Computing 5: 155-200.

Schmidt, K. and Simone, C., 2000. *Mind the Gap.* COOP 2000, Sophia Antipolis, France

Skok, W., 1999, *Knowledge management: London taxi cabs case study.* Special Interest group on Computer Personnel Research Annual Conference, Proceeding of the 1999 SIGCPR Conference on Computer Personnel Research, New Orleans, Louisiana, USA, ACM Press.

Sørensen, C. & Pica, D., 2005, *Tales from the Police: Mobile Technologies and Contexts of Work.* Information and Organization, vol. 15, no. 3.

Townsend, A., 2003, *Cabbie.* Stroud - UK, Sutton Publishing Limited.

Weilenmann, A., 2003, *Doing Mobility.* Gothenburg Studies in Informatics - PhD Thesis. Gothenburg, Gothenburg University. Sweden: 182.

Wiberg, M., 2001, *In between Mobile Meetings: Exploring seamless ongoing interaction support for mobile CSCW.* PhD Dissertation. Department for Informatics, Umeå University.

Wiredu, G., 2005, *Mobile Computing in Work-Integrated Learning: Problems of Remotely Distributed Activities and Technology Use.* London School of Economics PhD Thesis.

ARCHITECTURE FOR MULTI-CHANNEL ENTERPRISE RESOURCE PLANNING SYSTEM

Karl Kurbel, Anna Maria Jankowska, and Andrzej Dabkowski
European University Viadrina, Chair of Business Informatics
POB 17 86, 15207 Frankfurt (Oder), Germany

Abstract: Mobile computing is changing the behavior of individuals and organizations. Instant, multimodal access to information is beneficial in many business situations. Consequently, core information systems like Enterprise Resource Planning systems that today's organizations rely on have to support the mobile behavior of their users. In this paper we discuss some architectural considerations for multi-channel applications and introduce a four-tier architecture for a mobile ERP system. Two key questions to answer are how to access content of an ERP database from heterogeneous mobile devices, and how to make that content available in different formats to a mobile user. A prototypical implementation based on a real ERP system is described. Open questions and issues for further research are discussed in the concluding section.

Keywords: mobile computing, ERP system, multi-tier architecture, multi-channel applications, graphical and vocal user interfaces.

1. INTRODUCTION

An increasingly important requirement for the core information systems in enterprises is to provide support for the mobile behavior of their users. This trend goes hand in hand with ubiquitous computing (Weiser, 1991), i.e. guaranteeing access to information and computing power independent of locations and devices.

Network technologies for mobile business are maturing, becoming more and more powerful. With the introduction of third generation networks like UMTS (Universal Mobile Telecommunication System) with transfer rates up to 2 Mbps the limitations imposed by narrow bandwidths are relaxed. In

Japan, NTT DoCoMo's third generation i-mode service was launched in 1999 already (Yamakami, 2002).

In the long run, it can be expected that mobile devices will provide similar user interfaces as desktop monitors. This trend raises new challenges for business information systems in general and for enterprise resource planning (ERP) in particular. A long-term vision for ERP systems is to make all functionality available independent of particular front-end devices – on mobile high-resolution multimedia phones as well as on traditional desktop clients.

Our first step towards this vision is to make ERP data available for mobile users. Such data are normally stored in the ERP system's database and managed by a database management system (DBMS). The core questions are thus how to transmit queries of the mobile user from the mobile device to the DBMS used by the Enterprise Resource Planning system, and how to transmit and convert such data from the database tables so that they can be displayed on the screen of a mobile device. The two major aspects of a solution are:

1. Accessing the content of the database.
2. Extracting information retrieved from the database and preparing it in a device-dependent manner.

These two tasks are solved in our approach by a *Content Access Engine with Cache Storage Structures* and a *Content Extraction Engine*. In the subsequent sections, an architecture around these two engines is presented. The Content Access Engine (CAE) is in charge of retrieving data from a relational database and representing them in an XML (W3C, 2004) format. The responsibility of the Content Extraction Engine is to detect the type of the user's device and to generate device-specific forms of the XML data in the respective markup language for the user's display.

This paper is organized as follows. In the next section, the general architecture for mobile ERP, the underlying concepts, and the technologies used are presented. Section 3 illustrates by means of a specific ERP system how this architecture was implemented in a particular case. Some observations and open questions for further research are discussed in the final section.

2. ARCHITECTURAL DESIGN OF MOBILE ERP

2.1 General Considerations for Mobile Applications

Typical application systems today have three major layers: presentation layer, business or application logic layer, and services layer (Britton, 2000,

pp. 91-106). The presentation layer provides the user interface and is responsible for the interaction between the user and the device. The application or business logic layer contains the business rules that drive the given enterprise. The services layer provides general services needed by the other layers, usually including database services, file services, print services, and communication services.

The functionalities of these three layers can be assigned to logical entities called *tiers*. Mobile applications are typically deployed with three-tier or multi-tier architectures. Such architectures allow for parallel development of tiers by application specialists and provides flexible resource allocation. They require more planning but reduce development and maintenance costs over the long term by leveraging code re-use and elasticity in product migration (Myerson, 2002).

In our work the need to develop an architecture arose from the fact that we had to find an effective way to make ERP system data and functionality available on mobile devices. The major technical requirement for mobile access to an ERP system is presentation of information in multiple formats. Wireless devices are equipped with different browsers that support various media formats. It is therefore necessary to deliver the content in different markup languages such as WML (WAP Forum, 2002), XHTML (W3C, 2003a) or HTML (W3C, 1999). An appropriate architecture should make it easy to add new formats, without changing the existing structure. In addition, many mobile devices are not only equipped with a browser but also support J2ME (Java 2 Platform, Micro Edition). This technology offers better graphical user interfaces than WML or XHTML (Hemphill & White, 2002).

Due to recent advances in digital speech processing technologies and the emergence of new, non-proprietary standards such as Voice Extensible Markup Language (VoiceXML) (W3C, 2003) and Speech Application Language Tags (SALT) (SALT Forum, 2002), it is now possible to enhance mobile applications with voice user interfaces (VUIs) based on speech recognition and synthesized voice output. Although ERP systems are not traditional telephony-based services they can also benefit from additional voice-enabled interfaces. VUIs can be deployed for retrieving information from a database or manipulating data in the database. Voice input could be applied for entering new data or editing existing data. Voice-enabled interfaces can enhance alternative access methods and allow users to interact with an application in a variety of ways, using speech, keyboard, stylus, etc. (so-called multimodal access). Each of these modes can be used independently or concurrently.

Our architecture for mobile applications is designed for thin-client (browser-based) and fat-client (J2ME) applications. In this architecture, ERP

system functionality can be accessed through mobile and wireless devices. The ERP system as such remains unchanged.

The architecture is divided into four tiers. The first tier, the data tier, is represented by the ERP system's database. The second tier has the specific application logic of the "mobilization" task encapsulated in the Content Access Engine with Cache Storage and RFC Server. Application logic is defined as the processes which "do the work" such as requesting data, returning data, formatting data, etc., for example building queries from a mobile user's request for information and preparing the results for processing.

The Content Access Engine transforms the data retrieved into XML format. It takes into account the device's characteristics and manages the dispatching of information retrieved in portions. The entire result set is kept and managed in user-specific cache structures on the application server. The amount of data that can be served in one portion depends on the display of a particular device. Special data formats were developed to simplify the process of XML generation.

A Remote Function Call (RFC) Server is used so that ERP functions can be invoked by remote clients. If a mobile user wants to add or modify ERP data, appropriate functions of the ERP system are activated.

The third tier has the challenging task of device-context aware content delivery to the user, incorporating the presentation logic in the Content Extraction Engine. This engine determines the type of the browser and the most important device characteristics, and then tailors the content to significant features of the device. The Content Extraction Engine implements the presentation logic (Paragon, 2003).

The Content Access Engine and the Content Extraction Engine are placed on the Tomcat 5 application server (Apache, 2005a). Tomcat offers clustering and load balancing capabilities that are essential for deploying scalable and highly available Web applications. To guarantee such application features we applied the vertical scaling type of clustering with four instances of Tomcat running on a single machine. In the future we plan to extend our solution with horizontal scaling clustering capability (Apache, 2005).

The fourth tier consists of different mobile and wireless devices like WAP-enabled cellular phones, PDAs, Palmtops, and Pocket PCs with their respective browsers and GUIs. J2ME applications and vocal applications serve as additional user interfaces.

2.2 Content Access Engine and Caching Mechanism

The general architecture underlying our discussion is outlined in Figure 1. The Content Access Engine operates on database tables, as data of an ERP system are usually stored and maintained by a database management system (DBMS). For the interaction between an ERP system and a DBMS two solutions are commonly used. The ERP system can operate directly on data maintained by the DBMS, updating, inserting or deleting them. Alternatively, the most frequently requested or the last requested operational data are stored in memory and accessed there. If the needed data is not in the cache it is retrieved from the database upon the first request. The cached data are periodically exchanged with data in the database tables.

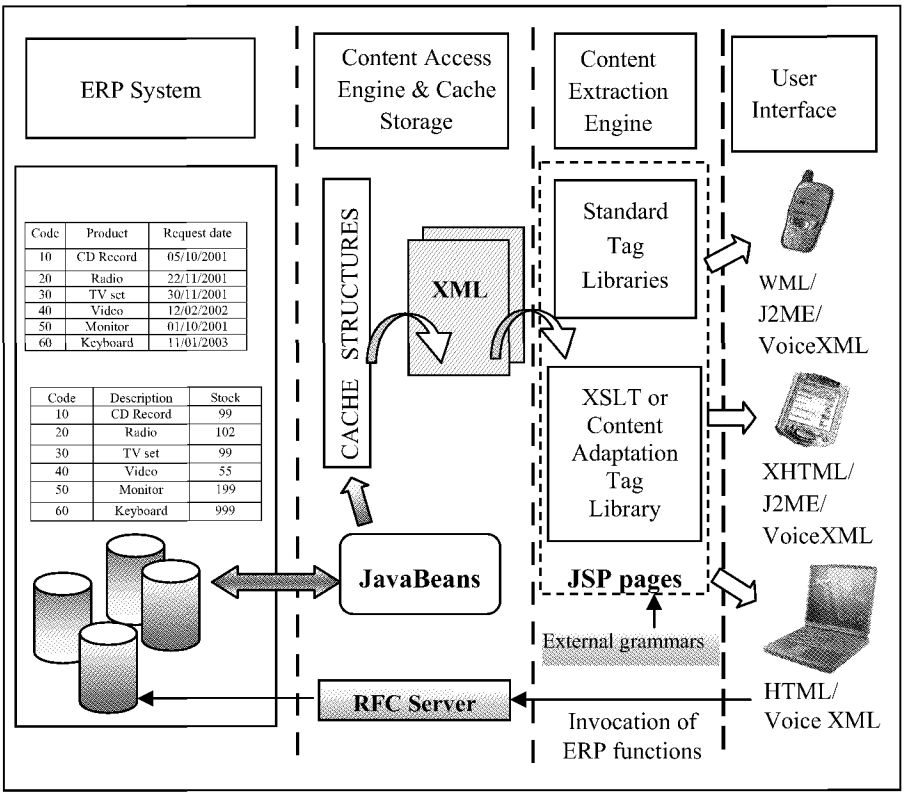


Figure 1. Architecture for mobile ERP

A cache structure is an integrated component of the Content Access Engine (CAE). It speeds up data access and improves the overall performance. The cache is designed to support and maintain the representation of result data. It is initiated with a list of results whenever a

user sends a request for new data. The cache structure belongs to a single user and it is accessible only to him/her. The cache is assigned to the user's session. It is freed when the user disconnects from the mobile ERP system or when the session expires.

Portions of data are served from the cache. Separate sets of parameters describing the devices' displays are assigned to different browsers and device types in the Content Extraction Engine. Subsequently, the parameters are passed to the cache when a browser communicates with the ERP system requesting data.

The task of retrieving data from a DBMS is accomplished by special components implemented as JavaBeans (Sun Microsystems, 1997). JavaBeans encapsulate all SQL queries that reflect the physical structure of the requested data in a database (Gertz, 2000). With regard to maintainability and portability of the CAE, JavaBeans are the components that have to be modified if the data organization, tables, or dependencies in the database change.

A reasonable way to store information from the database for further processing are dynamically built lists (in Java, vectors of vectors). The first row of such a list contains the column names from the database tables. The column names are used as tag names in the next step – the generation of XML documents on-the-fly based on the data in the list. However, when the XML documents are created, it is possible to replace the original column names with user-defined names. In this way a developer can provide more meaningful tag names.

2.3 Content Extraction Engine

A key requirement for mobile applications is to adapt content received from the ERP system to the characteristics of a specific device. In our approach this functionality is provided by the Content Extraction Engine. The Content Extraction Engine retrieves information about the devices' features and performs appropriate metadata transformations.

Some mobile operators use special proxies to adapt the content to the properties of mobile devices. Content transformation may also be performed directly on a client with the help of a mobile browser. Opera provides both solutions – it offers a Mobile Accelerator proxy and a special browser for handsets (Opera, 2005). The browser uses a set of style sheets to squeeze the content to the size of a mobile screen and scales images appropriately. The Mobile Accelerator reduces the page size, compresses the images and eliminates unnecessary content (e.g. banners). Both solutions yield pages tailored to a mobile screen but pagination or transformation of the site structure is not possible. Other known transcoding proxies (e.g.

WebAlchemist or Web Intermediaries) can perform advanced transformations based on transcoding heuristics and annotations, such as modifications of the site structure and of navigation modalities (Hwang, Seo and Kim, 2003; Hori et al., 2000). Special browsers and transcoding proxies, however, can provide only limited support in our solution since we already make content available tailored to the specific mobile device. Transcoding approaches or browsers may modify the content but only minor changes are left.

The most popular way to obtain delivery context is to use the HTTP standard Accept headers (W3C, 2002). These headers include the supported media types (MIME types), character sets, content encoding, and languages. Additionally, the User-Agent header contains information about the device manufacturer, the version number, the hardware, and the browser used. Integration of comprehensive context models can be achieved by using standards for contextual information like the CC/PP model developed by W3C (W3C, 2000) or UAProf introduced by the WAP Forum (WAP Forum, 2001).

In our approach content adaptation is based on information about device capabilities retrieved from HTTP headers or CC/PP profiles, if they are available. The Content Extraction Engine determines the form of presentation depending on the relevant features of a device – supported markup language, graphical formats, size of the display area, browser type, and colors displayed.

In the next step, the metadata generated by the Content Access Engine are transformed according to device-specific characteristics. Two components are available for this transformation: tag libraries and transformation objects with XSLT stylesheets (W3C, 2003b).

Tag libraries are reusable modules that can build and access programming language objects and influence the output stream. They usually encapsulate frequent tasks and can be used across applications, increasing the speed and quality of development. Tag libraries have access to all objects available to Java Server Pages, they can communicate with each other and can be nested, allowing for complex interactions within a page (Sun Microsystems, 2002).

Although some simple JSP tag libraries for mobile applications have already been provided by vendors, libraries supporting the development of applications for different devices are not yet available. Therefore we developed a special tag library – the Content Adaptation Tag Library – for the generation of appropriate markup elements depending on the device context. This library helps to separate the presentation format from the presentation logic. It encapsulates most of the functionalities used in HTML, WML, and XHTML pages.

The second component of the Content Extraction Engine are transformation objects with XSLT stylesheets. XML data can be transformed with XSLT into virtually any format. The most popular formats are WML, HTML and XHTML. For the transformations, it is necessary to prepare a number of stylesheets. An XML document then can be parsed and modified according to the respective stylesheet. In our approach the Java Transformation API for XML (TrAX) (Pfeifer, 2001) was chosen to invoke XSLT stylesheets from Java programs. TrAX is capable of compiling stylesheets and holding them in memory, thus improving the performance significantly. Up to now we have built a library of reusable stylesheets for generating HTML, WML, XHTML, VoiceXML and formatting XML to a special format for J2ME applications.

2.4 J2ME Front-ends

Presenting content in markup languages such as WML or HTML reduces the functionality of the handheld device to not more than a simple browser. The user interface is limited by the available markup elements, and when the user is not connected the browser is useless. Powerful mobile applications need user interfaces similar to those of desktop computers and functionalities beyond the scope of mobile browsers.

The Java 2 Platform, Micro Edition (J2ME) meets these requirements at least to some extent. J2ME was designed to accommodate a variety of embedded and hand-held devices. It is composed of both a configuration and a profile. A configuration consists of a virtual machine, core libraries, classes, and APIs. The configuration layer defines the minimum set of Java virtual machine features and Java class libraries available on a particular category of devices. The profile defines the minimum set of APIs available for a particular family of devices representing a given vertical market segment. Profiles are implemented on a particular configuration, and applications are written for a particular profile.

The Mobile Information Device Profile (MIDP) provides a set of Java APIs specific to a particular category of devices (cell phones, PDAs, etc.). In J2ME data can be cached on the client with the help of the MIDP Record Management Store (RMS) API and sent to a database when a connection is established (Riggs, Taivalsaari and VandenBrink, 2001). This solution is very helpful when the connection is suddenly interrupted. It can be applied for incoming and outgoing connections. Instead of temporarily caching data on the server, the RMS stores them on the mobile device and can work independently of the network connection. The same is true for entered data – they are first saved into the RMS and then sent to the database if a connection is available.

J2ME-based applications cannot directly connect to a database using the JDBC mechanism – therefore a special approach for the communication with a data source taking existing tiers into account is needed. Since J2ME technology does not provide any APIs for data connectivity, midlets have to be used to connect to a web server and get the data. Some kind of middleware (e.g. in form of JSP) is necessary for the communication with a database. In our architecture the Content Access Engine is responsible for data delivery in browser-based and J2ME applications. Since the data are represented in XML format, they can be used in J2ME applications, too. The entire structure of the CAE remains unchanged.

A mobile application based on the Java programming language organizes its graphical user interface as several forms in one file (midlet). Therefore the data in XML format cannot be processed in the same way as in an application based on a mobile browser. J2ME applications still need to manage operational memory and storage efficiently. From this point of view it is better to parse XML-based data somewhere else than on a mobile client. Keeping in mind these constraints, the Content Extraction Engine (CEE) can still be applied because JavaServer Pages can be used as a middleware between a MIDP front-end and XML data.

The Content Extraction Engine detects if a response is for a J2ME application. Then it transforms the data from XML format as created by the Content Access Engine to textual information, using either special XSLT stylesheets or tag libraries. Each line of text contains several fields with information items separated by spaces. A single line can be treated as one record in a database. The fields can be simply divided into tokens and displayed in a J2ME-based application on a form. If the set of results includes many records the textual sheet is made up of more than one line. In such a case all records, or a certain part of the records, are stored on the mobile device with the help of RMS.

2.5 Vocal Interfaces

Currently customer services can be provided via the Web and the telephone. In order to obtain telephone services the user needs a device that has audio interaction capabilities. The customer has to call an interactive voice response (IVR) platform that possesses audio input, output, telephony functions and its own service logic as well as a transaction server interface. Companies working on IVR systems developed their own markup languages for their applications, but customers were reluctant to invest in a proprietary technology. In 1999, AT&T, IBM, Lucent Technologies, and Motorola formed the VoiceXML Forum to establish and promote a new, non-proprietary standard based on the eXtensible Markup Language (XML) -

Voice eXtensible Markup Language (VoiceXML). VoiceXML is a language that has features to control audio output, audio input, presentation logic, call flow, telephony connections, and event handling for errors. It serves as a standard for the development of powerful speech-driven interactive applications accessible from any phone. With VoiceXML application developers do not have to care about issues such as concurrent threads of control, resource provisioning, and platform-specific APIs. The language also accommodates platform differences for services that place less importance on portability than on utilizing a new speech technology provided by an individual vendor.

Aural interfaces to the mobile ERP system are based on the VoiceXML standard. For static content (e.g. help information) digitized audio output (recorded voice) in form of mono 8KHz 8-bit u-Law .au-files was used. This format was chosen for performance reasons – files do not need as much storage and download time as the 16-bit linear files (IBM 2003, pp. 57-58). Synthesized speech was applied for the content retrieved dynamically from a database. The data extracted from a database are in the same XML-format as in the previous cases. They were transformed to the VoiceXML format using predefined XSLT style sheets. The Content Adaptation Tag Library was not used for the transformation due to the high complexity of the output structures produced. This degree of complexity was not encountered in the cases of the other formats. Grammars are generated dynamically from a database. For contextual help simple external grammars are utilized. For the design and test phases of the aural user interfaces, the IBM Voice Server Development Kit (IBM, 2002) which fully supports the VoiceXML standard was applied.

2.6 RFC Server

Reading ERP data and displaying them on a mobile appliance is important but not enough. Even more important is to make ERP functionality available to the mobile user. Complete integration of mobile devices with back-office systems is one of the targets for successful implementation of mobile solutions (Vaidyanathan, 2001).

Making ERP application functions available on a mobile device is a challenging task. It is more difficult than just accessing data in relational database tables. A function like re-scheduling a production job, for example, is much more complex than reading or updating values in the database. It may require not only modifications of data items but invoking other functions as well (Dreibelbis, Lacy-Thompson, 2000).

A *Remote Function Call (RFC) Server* is deployed in our architecture to trigger functions of the ERP system. This RFC Server is a real-time link

between an ERP system and a mobile device. It plays the role of a high level environment in which functions of the ERP system can be invoked remotely in real time (Narayanan, 2002). An RFC Server is tied to a particular ERP system. It encapsulates all specific details and communicates with the low-level API (Application Programming Interface) of that system.

When an RFC Server is provided developers of mobile applications do not need to bother with these low-level APIs to call an ERP function nor with specific implementation languages used by different vendors (ABAP for SAP R/3, PL/SQL for Oracle Financials, Lj4 for infor:COM, etc.). Data entered by a mobile user are taken as parameters for the business function to be called. For example, the user will say that he/she wishes to re-schedule a particular job, and enter the job number, the new delivery date, etc. on his device. These values are given as parameters to the remote function provided by the RFC Server. More precisely, a request with those parameters is generated and sent to a web server. The web server communicates directly with the RFC Server and invokes the appropriate business function. While it is not very difficult to develop suitable front-ends for such tasks, implementation of the remote functions invoking internal functions of the ERP system can be quite complicated.

Another problem is posed by the needed support for distributed ACID (Atomicity, Consistency, Isolation, and Durability) transactions. In distributed multi-user applications sharing objects in real time can lead to resource sharing conflicts (e.g. many users may want to update the same objects at the same time). To avoid such conflicts various locking strategies to preserve the integrity of changes are used. In our framework, so-called optimistic locking (write locking) is applied. Optimistic locking allows unlimited read access to an object but a client can only write an object to the database if the object has not changed since the client last read it (Adya, 1994). If many users have the same data open, only the first update to commit succeeds while the others obtain error messages.

If the mobile device sends a request to an application server to update some data in the database, the application server invokes an appropriate ERP function via the RFC server. If that function cannot be successfully executed, an exception (`OptimisticLockException`) is thrown, propagated further to the RFC server and subsequently to the application server. The application server sends a response to the mobile device informing about the error and the user is asked to refresh the current data and re-enter values. Write-write conflicts in our solution are detected using timestamps indicating the last commit time.

3. AN EXAMPLE OF MOBILIZING A REAL-WORLD ERP SYSTEM

3.1 Restrictions and Design Considerations

Significant problems in mobilizing today's ERP systems are caused by low bandwidths of telecommunication networks. It is time and cost ineffective to send larger portions of data to mobile devices. Processing should better be done on an application server and only the results should be transmitted in a compact form to the mobile device.

Another shortcoming is the low processing power of mobile devices. As a consequence, only simple things like validating input can be done directly on the device while the mobile application logic must remain on an application server. We took these aspects into account when developing a prototypical solution for a real-world ERP system, *infor:COM* by Infor Global Solutions (Infor, 2005). This system aims at small and medium-size enterprises and has a good market penetration in Central Europe.

Before selecting those ERP application domains which appeared worth to be enhanced with mobile access, an empirical study of the state-of-the-art in this field was done (Kurbel, Teuteberg, Hilker, 2003). Then the infor:COM system was analyzed with respect to application areas which are both interesting from a business point of view and feasible taking technical limitations into account.

3.2 Content Access and Content Extraction

Standard infor:COM applications have a forms oriented user interface, spreading information all over a conventional screen of a large monitor. Screen content can be quite complex, showing many data and sub-forms at the same time.

Considering the small display sizes of mobile devices, it is obviously not possible to "translate" an existing ERP front-end into a mobile front-end one-to-one. The front-end designer has to concentrate on important information instead and adopt only the really essential data from the standard screens of the desktop-based application. Usually it is necessary to divide the data up into several screens of a mobile device ("cards" in the terminology of mobile browsers).

As an example, processing of quotes in the sales module of infor:COM is described. On a standard desktop client, the user generally sees a large form with many menus, submenus, drop-down-boxes, text fields, etc. It is quite obvious that this type of user interface has to be re-designed because it cannot be displayed in the same form on the screen of a mobile device.

Therefore the menus in our mobile ERP application are displayed level by level. When the user starts navigating on the top level and clicks on a menu item, he or she is re-directed to menus detailing the functions of the chosen module. Finally the user reaches a card with the desired functionality.

Figure 2 shows a search for quotes that match some conditions specified by the user as a stepwise process on a mobile device. The front-end was implemented with the help of a Nokia 7210 simulator for mobile devices. Such simulators are available in toolkits provided by devices manufacturers and other vendors (Kurbel, Dabkowski, Zajac, 2002).

Figure 2. Selecting quotes from ERP database

Assuming that the user looks for a particular quote (quote no. "AG1001"), he or she navigates through a sequence of menus to reach the desired card. The functionality outlined in Figure 2 is the same as for the desktop-based infor:COM system. For example, the user can select a certain way of filtering quotes (by quote number, RFQ number, customer number, etc.). From the list of quotes displayed afterwards he or she can select the desired one. If the list of results is long, only a few items are displayed at a time, i.e. the list is distributed over several cards. Likewise the details of a quote are also divided into a number of cards.

While the user sees only the front-end displaying ERP data as requested, the requests and responses are transferred across the overall architecture as described in the previous section. Looking for information the user fills input fields on a mobile screen. In this way he or she determines the criteria for the

search. The user's input is treated as parameters and passed to the Content Access Engine where it is further processed behind the curtain.

4. OPEN ISSUES AND FURTHER WORK

Industry experts forecast that the market for wireless applications will continuously grow over the next few years. With increasing needs of business users to access information from anywhere at any time, organizations have to find new, more effective system architectures.

This paper focused on building a multi-tier framework for mobilizing an existing ERP system. An alternative approach for a mobile ERP solution could be based on a Web Services architecture (Chappell and Jewell, 2002). Web Services are considered as a major step forward in inter-enterprise cooperation and integration of different types of systems. They can play the role of a universal Application Programming Interface which does not require any other protocol than the Internet protocols. Web Services standardize the calling, exchange and organization of application services. In terms of software development, the increasing complexity of business information systems and the need for rapid adaptation to different devices and for integration of business concepts are good reasons for using a high-level approach which implements techniques of modeling and programming with business objects. In our project reengineering and development work is currently going on. The goal is to implement an effective, Web Services-based architecture for a part of the mobile ERP system.

References

Adya, A., 1994, Transaction Management for Mobile Objects using Optimistic Concurrency Control; http://research.microsoft.com/~adya/pubs/tr.pdf.

Apache, 2005, Clustering/Session Replication How-To; http://jakarta.apache.org/tomcat/tomcat-5.0-doc/cluster-howto.html.

Apache, 2005a, Tomcat Server; http://jakarta.apache.org/tomcat/.

Britton, Ch., 2000, *IT Architectures and Middleware: Strategies for Building Large, Integrated Systems*, Addison-Wesley, Boston.

Chappell, D. and Jewell, T., 2002, *Java Web Services*, O'Reilly & Associates, Sebastopol.

Dreibelbis, D. and Lacy-Thompson, T., 2000, Interfacing with SAP R3; http://www.eaijournal.com/Article.asp? ArticleID=144&DepartmentId=4.

Gertz, M., 2000, Oracle/SQL Tutorial; http://www.db.cs.ucdavis.edu/teaching/sqltutorial.

Hemphill, D. and White, J., 2002, *Java 2 Micro Edition*, Manning Publications, Greenwich.

Hori, M. et al., 2000, Annotation-Based Web Content Transcoding, in: Proceedings of the 9th International World Wide Web Conference on Computer Networks, Amsterdam, pp. 197-211.

Hwang, Y., Seo, E., Kim, J., 2003, Structure-Aware Web Transcoding for Mobile Devices, in: *IEEE Internet Computing* 7 (5): 14-21.

IBM, 2002, IBM: Voice Server SDK; http://www-3.ibm.com/software/voice/.

IBM, 2003, VoiceXML Programmer's Guide; http://www.elink.ibmlink.ibm.com/public/applications/publications/cgi bin/pbi.cgi.

Infor Global Solutions, 2005, infor:COM; http://www.infor.de.

Kurbel, K., Dabkowski, A. and Zajac, P., 2002, Software Technology for WAP-based M-Commerce - a Comparative Study of Toolkits for the Development of Mobile Applications, in: Proceedings of the International Conference WWW/Internet 2002 (IADIS), Lisbon, pp. 673-676.

Kurbel, K., Teuteberg, F. and Hilker, J., 2003, Mobile Business-Anwendungen im Enterprise Resource Planning: Mobilitätspotentiale entlang der ERP-Funktionskreise, *Industrie Management* 19 (1): 72-75.

Myerson, J. M., 2002, *The Complete Book of Middleware*, Auerbach Publishers, Philadelphia.

Narayanan, V., 2002, Interfacing with SAP R/3; http://www.info-sun.com/docs/wp_sapinter.pdf.

Opera, 2005, Opera Products for Mobile; http://www.opera.com/products/mobile/.

Paragon Corporation, 2003, Separation of Business Logic from Presentation Logic in Web Applications; http://www.paragoncorporation.com/ArticleDetail.aspx?ArticleID=21.

Pfeifer, C., 2001, XML Processing with TraX; http://www.onjava.com/pub/a/onjava/2001/07/02/trax.html.

Riggs, R., Taivalsaari, A. and VandenBrink, M., 2001, Programming Wireless Devices with the Java 2 Platform Micro Edition, Addison-Wesley, Boston.

SALT Forum, 2002, Speech Application Language Tags 1.0 Specification; http://www.saltforum.org/saltforum/ downloads/SALT1.0.pdf.

Sun Microsystems, 1997, JavaBeans; http://java.sun.com/products/javabeans/docs/spec.html.

Sun Microsystems, 2002, JavaServer Pages Specification, Version 2.0; http://jcp.org/aboutJava/communityprocess/first/ jsr152/.

Vaidyanathan, R., 2001, Wireless Application Integration. EAI Journal; http://www.eaijournal.com/Article.asp? ArticleID=450&DepartmentId=3.

W3C, 1999, HTML 4.01 Specification. 1999; http://www.w3.org/TR/html4/.

W3C, 2000, Composite Capabilities/Preference Profiles: Terminology and Abbreviations, Working Draft; http://www.w3.org/TR/2000/WD-CCPP-ta-20000721/.

W3C, 2002, Delivery Context Overview for Device Independence; http://www.w3c.org/2001/di/public/dco.

W3C, 2003, Voice Extensible Markup Language (VoiceXML) Version 2.0; http://www.w3.org/TR/voicexml20/.

W3C, 2003a, XHTML 2.0 The Extensible HyperText Markup Language Specification; http://www.w3.org/TR/xhtml2/.

W3C, 2003b, XSL Transformations (XSLT) Version 2.0; http://www.w3.org/TR/xslt20.

W3C, 2004, Extensible Markup Language (XML) 1.0 (Third Edition); http://www.w3.org/TR/REC-xml.

WAP Forum, 2001, UAProf; http://www1.wapforum.org/tech/terms.asp?doc=WAP-248-UAProf-20010530-p.pdf.

WAP Forum, 2002, Wireless Application Protocol WAP 2.0, Technical White Paper; http://www.wapforum.org/what/ WAPWhite_Paper1.pdf.

Weiser, M., 1991, The Computer for the 21st Century, *Scientific American* 265 (3): 94-104.

Yamakami, T., 2002, Leveraging Information Appliances: a Browser Architecture Perspective in the Mobile Multimedia Age, in: Chen, Yung-Chang et al. (Eds.): Proceedings of the 3rd IEEE Pacific Rim Conference on Multimedia (PCM 2002), Taiwan, pp. 1-8.

TOWARDS MOBILE INFORMATION SYSTEMS

Juhani Iivari
Department of Information Processing Science, University of Oulu, P.O. Box 3000, FIN-90014 Oulun yliopisto, Finland

Abstract: The contention of the present paper is that information systems proper form a significant application area of mobile technology. For that purpose the paper analyses the concept of mobility in the context of information systems using a sound metamodel for information systems. The anlysis makes it possible to make sense of different examples of mobile information systems and possibly to innovate new ones.

Keywords: mobile information system, mobile application, mobile service

1. INTRODUCTION

After the boom of mobile computing started in the early 1990's, the word "mobile" is easily used just as a selling argument without much attention to its meaning. It has served as a shorthand to mean that some service or product is accessed using a mobile device, a cellular phone for example. The term "mobile Internet" is an example of this. However, if we wish to have a more scientific understanding of mobility, we should be more careful with the terms.

Many experts claim that this first decade of the 21st century will be the decade of mobile computing (Urbaczewski et al., 2003), even though the expectations tend to be unrealistically high as in the case of every technology boom. In fact, the mobile field still waits for a second "killer application", if short messages are considered the first. It may be that it is not realistic to expect new universal or horizontal killer applications anymore. Instead the applications of mobile computing may be more vertical targeted to

specific application domains. This paper considers information systems one possible application area of mobile computing.

Accordingly, the present paper focuses on mobility in the context of information systems. It analyses different aspects of spatial mobility in the IS context. To ensure systematic analysis and to integrate the aspects of mobility, the paper applies a sound metamodel for an information system (Iivari, 1989; Iivari 1990) as a structuring framework. The analysis provides a framework for understanding mobility, to innovate new mobile applications, and to develop them.

2. THE METAMODEL FOR AN INFORMATION SYSTEM

An information system (IS) is a computer-supported system with the purpose "to supply its groups of users (...) with information about a set of topics to support their activities" (Gustafsson et al., 1982). In addition to conventional information systems the above definition covers all digital content that aims at informing about some topic (such as company web pages).

In consistence with Gustafsson et al. (1992), Iivari (1989, 1990) proposes a metamodel for an information system consisting of three major levels of modeling or abstraction for an information system:

1. The organizational level (the users and their activities).
2. The conceptual/infological level (information and the topics).
3. The datalogical/technical level (the technical implementation)

The three levels of modeling are abstractions from three different, even though interdependent domains: the host organization, in which the information system is to be "embedded", the universe of discourse, about which the information system is to communicate information, and the technology to be used in the technical implementation of the system.

The organizational level consists of two sublevels: the (re)designed organizational context and the application concept. The (re)designed organizational context describes organizational actors/units, organizational functions ("business processes") affected and/or redesigned by the information system in question. The application concept summarizes the IS services - organizational information types, information processing functions and communication relationships - delegated to the information system.

The metamodel for the conceptual/infological level identifies three sublevels: the Universe of Discourse (UoD), IS specifications and the user interface. The idea of the UoD level is to formalize the model of the "real" world underlying the information system. The level of IS specifications describes the information content and functionality of the system, ignoring the external

representation of the system to the users. The external representation is the topic to be addressed at the level of user interface. The user interface defines how the content abstractions are made accessible and visible to the user.

Information systems are information specific, they inform about specific topics. An essential part of the systems development is to specify their information content and structure. The above metamodel provides and integrates three perspectives into the analysis and design of the information content of system: the "ontological-conceptual/" perspective, the "information-document perspective" and the "user interface" perspective.

The above metamodel has successfully been applied to identify and categorize objects in the object-orientation analysis (Iivari, 1990) and to analyse object-oriented methods (Iivari, 1994) as well as to structure IS security issues (Siponen, 2001). In the following it will be applied to analyse mobility in the context of information systems.

3. MOBILITY IN THE CONTEXT OF INFORMATION SYSTEMS

The metamodel for an information system allows to integrate a number of aspects of mobility into a systematic framework. At the organizational level, and more specifically at the sublevel of (re)designed organizational (or interorganizational) context, one can distinguish business models for mobile services (especially for mobile commerce) and models of mobile work. Business models discussed in Tsalgatidou and Pitoura (2001) and Varshney (2003), for example, describe business actors, their roles and the product, service and information flows between them in providing the services. This paper will choose more a customer focus on the mobile activities/work to be supported by the mobile services. If a customer or a user is not specially interested in the underlying business model of a mobile service, the business model should be as invisible as possible to her.

Mobility of work means that the work may be performed at different places either locally (e.g. within a building) or more remotely. This mobility easily implies that the context of work may change. Gorlenko and Merrick (2003) distinguish mobile office context and mobile fieldwork context in the case mobile work. In the former case they assume that the computer is the primary task for the user and walking and manipulating other objects secondary. In the mobile field context moving and working with other objects other than the computer is primary, and the use of computer only secondary. Referring to Kristoffersen and Ljungberg (1999) they suggest that in the mobile field context visual attention of users is mainly directed to events outside the computer, users hands are commonly engaged with a variety of

physical objects unrelated with the computer device, and users often must be move all the time. Following again Kristoffersen and Ljungberg (1999) they propose that there are two ways of operating the computer in the mobile field context: the user may "make place" for interacting with the device, that is interrupt the main task in order to interact with the computer device, or "take place", i.e. to interact with the device while executing the main task (e.g. repair work on the field).

The study of Perry et al. (2001) mainly concerns mobile office context. They followed closely business trips of 17 mobile workers who mainly worked in multiple, but stationary locations. They summarize their findings in four points: planning opportunistic access to information (e.g. carrying particular technologies and documents just in case), working in dead time, use of the mobile phone as a device proxy (accessing remote technological and informational resources using a mobile phone), and use of technologies for remote awareness monitoring (of activities of remote colleagues).

Table 1 suggests that changing context is not mobility as such, but a natural consequence of mobility. Spatial mobility implies that the location of the user and the device may change, and consequently the physical environment (e.g. temperature and light), artifacts in the environment (especially ICT artifacts capable to communicate with the focal information system in question), users activities may also change while moving, and participants involved.

The distinction between wandering, traveling and visiting (Dahlbom and Ljungberg, 1998) clearly concerns potential users of the system. Wandering users rather than traveling or visiting ones may be more interesting in the case of mobility, since the latter can easily be supported by portable devices such as laptops. One should note, however, these are not mutually exclusive categories. A traveling or visiting user may be wandering at the same time and even in the same situation (e.g. wandering in a train while traveling). Then a crucial question often is if one device can satisfactorily support wandering, traveling and visiting.

The distinction between on-site movers, yo-yos, pendulums, nomads, carriers (Lilischkis, 2003) can be interpreted as a hybrid one that tells something about the mobile work and about potential users of the systems developed to support the mobile work in question.

Table 1. Dimensions of mobility of mobile information systems

Level	Dimension of mobility
Organizational level	
(Re)designed organizational or inter-organizatio-	(Business models for mobile services) Mobile work/activity - local mobility, remote mobility (Luff and Heath, 1998) - mobile office context (Gorlenko and Merrick, 2003; Perry et al., 2001)

nal context	vs. mobile field context (Gorlenko and Merrick, 2003; Pascoe et al., 2000) - on-site movers, yo-yos, pendulums, nomads, carriers (Lilischkis, (2003) = > Changing context (≈ Gorlenko and Merrick, 2003; Krogstie et al., 2003; Tarasewich, 2003): - location, environmental context, artifact context, activity context, participant context - activity "mobility", social "mobility" (Lyytinen and Yoo, 2002)
Application concept	Mobile user - wandering, traveling, visiting (Dahlbom and Ljungberg, 1998) - on-site movers, yo-yos, pendulums, nomads, carriers (Lilischkis, (2003) Wireless interaction - stationary interaction (Perry et al., 2001) – mobile interaction (Gorlenko and Merrick, 2003; Pascoe et al., 2000) Context-aware IS services – non-context-aware IS services - location-aware, environment-aware, artifact-aware, activity-aware, participant- aware, user-aware (≈ Gorlenko and Merrick, 2003, Krogstie et al., 2003; Tarasewich, 2003)
Conceptual/infological level	
Universe of Discourse	Fixed, portable and mobile entities (Varshney, 2003)
IS specifications	Location-dependent information about mobile entities => Context-aware functionality
User interface	Mobile terminal Micro mobility (Luff and Heath, 1998) => Context-aware user interface => Display mobility
Datalogical/technical level	
Hardware and network	Mobile devices Wired – wireless
Data	Mobile data (data migration)
Programs	Mobile programs (program migration)

When considering the application concept for a mobile information system, one should decide whether stationary of mobile wireless interaction (Gorlenko and Merrick, 2003) with the system is needed. In the stationary wireless interaction the interaction session is assumed to take place in one place while in mobile interaction both the user and the device may change the location while interacting.

Another aspect of the application concept concerns whether the system is context aware or not. Mobility almost by definition implies that that the location of the user changes. It is not necessary to restrict context-awareness to location only (Schmidt et al., 1999).

When looking at mobility at the conceptual/infological level, the first aspect is that the entities of interest are mobile. The users may also be entities of interest in the system. If their location is of interest, it may be followed by GPS, for example. Because of the changing location of users location-sensitive services form a natural option in the case of mobile systems.

At the level of IS specifications one specifies the functionality of the system in detail. If it was decided at the level of the application concept that the application will be context-aware proving context-aware IS services, one must specify in detail how the provided functionality is made context-aware.

Mobile devices include user interface problems of their own that have been discussed, for example, by Pascoe et al. (2000) and York and Pendharkar (2004). In addition to the micro-mobility of the mobile devices one can imagine display mobility, i.e. the depending on the alternative display devices available in the user's current location, the system selects the most appropriate display device available (e.g. PDA, TV set) possibly depending on the user's current attention (Banavar and Bernstein, 2002).

The datalogical/technical level includes implementation issues such as software on mobile terminals, capabilities of mobile devices, middleware for mobile applications and wired and wireless networks (Varshney and Vetter, 2002). The capacity of the handheld devices is limited and the network connectivity may vary. From the viewpoint of the user the mobile device together with the connected network forms a distributed information/database system. The data required for the service may reside in different nodes of the network. Even though necessary and useful, the wireless access to World Wide Web is a limited solution, since it assumes that the processing takes place in predetermined nodes. In the mobile environment there is a natural need to have mobile data and mobile programs, which are able to migrate in the network depending on the access and availability of capacity (Pitoura et al., 2001). It is beyond the scope of the present paper to discuss these implementation details.

4. TOWARDS MOBILE INFORMATION SYSTEMS

The proposed framework can be applied in classifying mobile applications and potentially in innovating new ones. One can start with mobile user, noting that even traveling, visiting and/or wandering users (Kristoffersen and Ljungberg, 1999) do not form homogenous categories. Traveling in a train obviously differs from traveling in a car, if one also drives the car. Also from the point of use, traveling in a train with a seat is quite different from traveling without a seat. Also visiting differs depending on whether one has a convenient wired access to computers or not.

Fieldwork is an example of mobile users doing mobile activity. It has received considerable attention in the context of mobile computing. As York and Pendharkar (2004) point out field workers form a very diversified group.

Mobile entities form one natural application area for mobile information systems, if their location is of interest. The entities may be human beings, animals, vehicles or parcels.

One can identify more specific applications combining different aspects of mobility such as mobile users with mobile entities. Proxy Lady (Dahlberg et al., 2002) is an example of a mobile application supporting opportunistic communication of mobile users mainly in the work context. One can imagine an analogous mobile dating service in the context of big festivals, for example. Based on the user profile and preferences concerning the partner, the system might signal, if a suitable partner is in the vicinity.

The final case is mobile users doing mobile activity/work involving mobile entities of interest. One could speculate that this is the optimal context for mobile technology. However, this market niche may be quite restricted, since there is evidence that walking and working do not necessarily fit well together (Barnard et al, 2005). Walking tends to reduce the performance of work, since walking requires more attention than sitting or standing. Therefore it may be that mobile interaction is reasonable only when the nature of work is so inherently mobile that the user cannot interrupt its execution and make place and take time for the stationary interaction.

5. FINAL COMMENTS

This paper has attempted to analyse systematically the nature of and opportunities for mobile services and information systems. It has deliberately omitted various reasons to adopt mobile applications such as their impact on efficiency and effectiveness, customer and employee satisfaction, cost, and security (Nah et al., 2005). These are, of course pragmatically significant issues to be considered. Limited security and privacy of mobile computing and the danger of unwanted surveillance (Veijalainen and Visa, 2003) naturally lower customer trust in mobile applications (Siau and Shen, 2003) and slow down their adoption. The paper has also omitted the issue of how the development and delivery of mobile services can be organized in terms of content providers, application developers, service provides and network operators (Alahuhta, 2005; Varshney and Vetter 2002), with the idea that it is most important to consider first the system providing the services as a totality from the viewpoint of customers and users. It is the belief of the paper that the systematic understanding of the object of work, the mobile information system as the target of development, provides a useful starting point to consider of how to develop mobile applications and services.

The implications of the proposed metamodel for development can be summarized in the following seven points:

- Analyse the mobility of work and/or activities to be supported by the system.
- Analyse the nature of mobility of the intended users of the system.
- Consider whether mobile interaction with the system is required or whether wireless stationary interaction is sufficient.
- Consider whether context-awareness of the application is required or adds sufficiently value to users.
- Consider what are the entities of interest to the users. Are any of them mobile and is the location information about them required?
- Specify the functionality of the system. If the application is to be context-aware, specify the context-awareness of the system in detail.
- Design the user interfaces of the system. If the system includes mobile entities, consider how their location and movement information is captured (automatically or as human inputting). In the case of output (service) users, consider how to represent the functionality of the system to the user, design the user interfaces and the user interaction, taking into account the nature of work, the interaction mode (stationary vs. mobile), possible context-awareness (possibly requiring special sensors), the attention required by the work outside the mobile device, and the restrictions and capacity imposed by the mobile devices available for the implementation.

References

Alahuhta, P., Ahola. J. and Hakala, H., 2005, *Mobilizing Business Applications*, TEKES, Technology Review 167/2005, Helsinki

Banavar, G. and Bernstein, A., 2002, Software infrastructure and design challenges for ubiquitous computing applications, *Communications of the ACM*, **45**(12): 92-96

Barnard, L., Yi, J.S., Jacko, J.A. and Sears, A., (2005) An empirical comparison of use-in-motion evaluation scenarios for mobile computing devices, *International Journal of Human-Computer Studies*, (in press)

Dahlberg, P., Ljungberg, F. and Sanneblad, J., 2002, Proxy Lady: Mobile support for opportunistic interaction, *Scandinavian Journal of Information Systems* **14**(1): 3-17

Dahlbom, B. and Ljungberg. F., 1998, Mobile Informatics, *Scandinavian Journal of Information Systems*, **10**(1&2): 227-234

Gorlenko, L. and Merrick, R., 2003, No wires attached: Usability challenges in the connected mobile world, *IBM Systems Journal*, **42**(4): 639-651

Gustafsson, M.R., Karlsson, T. and Bubenko, J. Jr., 1982, A declarative approach to conceptual information modeling, in Olle, T.W., Sol, H.G. and Verrijn-Stuart, A.A. (eds.): *Information systems design methodologies: a comparative review*, North-Holland, Amsterdam: 93-142

Iivari, J., 1989, Levels of abstraction as a conceptual framework for an information system, in Falkenberg, E.D. and Lindgreen, P. (eds.), *Information Systems Concepts: An In-Depth Analysis*, North-Holland, Amsterdam: 323-352

Iivari J., 1990, Hierarchical spiral model for information system and software development, Part 1: theoretical background, *Information and Software Technology*, **32**(6): 386-399

Iivari J., 1991, Object-oriented information systems analysis: A framework for object identification, in Shriver B.D (ed.), *Proceedings of the Twenty-Fourth Annual Hawaii International Conference on Systems Sciences*, Vol. II, IEEE Computer Society Press: 205-218

Iivari, J., 1994, Object-oriented information systems analysis: A comparison of six object-oriented analysis methods, in Verrijn-Stuart, A.A. and Olle, T.W. (eds.), *Methods and Associated Tools for the Information Systems Life Cycle*, IFIP Transactions A-55, North-Holland: 85-110

Kristoffersen, S. and Ljungberg, F., 1999, "Making place" to make IT work: Empirical explorations of HCI for mobile CSCW, *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work*, ACM Press: 276-285

Krogstie, J., Lyytinen, K., Opdahl, A., Pernici, B. Siau, K. and Smolander K., 2003, Mobile information systems – Research challenges on the conceptual and logical levels, in Olivé A. (ed.), *ER 2003 Workshop*, LNCS 2784, Springer-Verlag, Berlin

Lilischkis, S., 2003, *More Yo-yos, Pendulums and Nomads:Trends of Mobile and Multi-location Work in the Information Society*, Issue Report 36, Star-project/Empirica.

Lyytinen, K. and Yoo, Y., 2002, Research commentary: The next wave of nomadic computing, *Information Systems Research*, **13**(4): 377-388

Nah, F.F-H., Siau, K. and Sheng, P., 2005, The value of mobile applications: A utility company study, *Communications of the ACM*, **48**(2): 85-90

Pascoe, J, Ryan, N. and Morse, D., 2000. Using while moving: HCI issues in fieldwork environments, *ACM Transactions on Computer-Human Interaction*, **7**(3): 417-437

Perry, M., O'Hara, K., Sellen, A., Brown, B. and Harper, R., 2001, Dealing with mobility: Understanding access anytime, anywhere, *ACM Transactions on Computer-Human Interaction*, **8**(4): 323-347

Pierre, S., 2001, Mobile computing and ubiquitous networking: concepts, technologies and challenges, *Telematics and Informatics*, **18**: 109-131

Pitoura, E. and Samaras, G., 2001, Locating objects in mobile computing, *IEEE Transactions on Knowledge and Data Engineering*, **13**(4): 571-591

Schmidt, A. Beigl, M. and Gellersen, H.-W., 1999, There is more to context than location, *Computer & Graphics*, **23**: 893-901

Siau, K. and Shen, Z., 2003, Building customer trust in mobile commerce, *Communications of the ACM*, **46**(4): 91-94

Siponen M.T., 2001, A Survey of the recent IS security development approaches: descriptive and prescriptive implications, in: Dhillon, G. (ed.), *Information Security Management - Global Challenges in the Next Millennium*, Idea Group Publications Hershey, PA: 101-124

Tarasewich, P., 2003, Designing mobile commerce applications, *Communications of the ACM*, **46**(12): 57-60

Tsalgatidou, A. and Pitoura, E., 2001, Business models and transactions in mobile electronic commerce: requirements and properties, *Computer Networks*, **37**: 221-236

Varshney, U., 2003, Location management for mobile commerce applications in wireless Internet environment, *ACM Transactions on Internet Technology*, **3**(3): 236-255

Varshney, U. and Vetter, R., 2002, Mobile commerce: Framework, applications and networking support, *Mobile Networks and Applications*, **7**: 185-198

Veijalainen, J. and Visa, A., 2003, Guest editorial: Security in mobile computing environments, *Mobile Networks and Applications*, **8**: 111-112

Urbaczewski, A., Valacich, J.S. and Jessup, L.M., 2003, Mobile commerce, Opportunities and challenges, *Communications of the ACM*, **46**(12): 31-32

York, J. and Pendharkar, P.C., 2004, Human-computer interaction issues for mobile computing in a variable work context, *International Journal of Human-Computer Studies*, **60**: [771-797

A MULTI-ACTOR, MULTI-CRITERIA APPROACH FOR TECHNOLOGY SELECTION WHEN DESIGNING MOBILE INFORMATION SYSTEMS

Jan Ondrus[1], Tung Bui[2], and Yves Pigneur[1]
[1]*University of Lausanne, Switzerland;* [2]*University of Hawaii at Manoa, USA*

Abstract: The fast ever-growing number of newly introduced mobile technologies makes the development of mobile information systems a somewhat complex activity. Decision makers - providers, merchants, and consumers alike - have to face great uncertainty and complexity regarding the acceptance of mobile technologies. Therefore, we stress that the selection process of an enabling technology for mobile commerce should be preceded with the use of a structured assessment methodology. With different available alternatives and various criteria for technology evaluation, multi-criteria decision making methods seem to be appropriate to support this selection process. Moreover, the success of introducing a new technology in a mobile information system depends on the preferences of varied involved actors in the market. We also consider in our approach the existence of multiple actors for the search of a technological consensus. As an illustration, we apply our approach to the mobile payment industry.

Keywords: Technology selection, multi-criteria decision making methods, Electre, mobile payments

1. INTRODUCTION

The widespread adoption of mobile devices has paved the way for the development of many innovative applications. However, the design of such applications or mobile information systems raises critical technical as well as business issues. This is partly due to the uncertainty surrounding the anticipated success or failure of enabling technologies. The traditional software and requirements engineering concepts, tools, and methods for

analyzing, designing, and implementing mobile information systems are the essential perspectives of this workshop. The choice of the appropriate enabling technologies is key and has to be considered during the design process.

For example, in the transportation industry, a mobile information system would integrate various mobile services. One of them is the mobile payment system. During its development phase, the IS designers have to consider different technologies with a great amount of uncertainty. They have the choice between card-based and phone-based solutions. Apart from the physical form of the device, they also have to appraise the applicability of the possible embedded technologies in each device. In other words, RFID contactless cards are limited to physical transactions whereas mobile phones using SMS enable both physical and remote transactions. Moreover, for comparison purposes, they have to conduct an evaluation of various technological aspects such as the cost, the ease of use, and the security.

Thus, there seems to be a need for a more structured approach to support these types of evaluations and decisions. We contend that multi-criteria decision making (MCDM) approaches are well adapted for the problem at hand. MCDM methods imply a modeling activity, which should clarify many aspects, making the decision process more transparent. Moreover, Stewart considers MCDM to be largely concerned with the deployment of systematic methods to help address problems characterized by incomparable objectives, multiple stakeholders and conflicting interests (Stewart, 1992). Consequently, MCDM methods seem to be appropriate for technology selection in a multi-actor context where technological consensus is vital for success.

The objective of the paper is to illustrate the feasibility of using MCDM methods to select enabling technologies in the design process of mobile information systems. This paper is structured as follows: in Section 2, we briefly discuss the potential of MCDM methods for technology assessment. In Section 3, we present a MCDM procedure for the technology selection, using a well-known technique of preferences aggregation with first insights into the mobile payment industry. Finally, in Section 4, we conclude with a summary.

2. MCDM FOR TECHNOLOGY ASSESSMENT

Multiple criteria decision methods, in general, have proven useful in supporting decision making (Keen, 1977; Zeleny, 1982). Few attempts were made to assess technologies with MCDM methods. Chan et. al propose to use fuzzy MCDM method to determine best technology selection. They

present a systematic approach using the concept of fuzzy set theory and hierarchical structure analysis to help decision makers to make suitable decisions in uncertain environments (Chan et. al, 2000). Salo et al. suggest the use of MCDM for technology foresight. They argue that there is a potential "in terms of lending rigor and transparency to foresight process" (Salo et al., 2003). Chou et al. tried to apply a multi-criteria decision making, such as AHP (Saaty, 1980), to assess mobile payment (Chou, et al., 2004). They analyze and explain the performance of different current payment instruments with technological, economic, and social factors. Their objective was to use these different factors in order to explain success or failures. One motivation of this analysis is that a payment system: "can be flawed technologically but still become the de facto e-payment scheme due to the advantage of an established customer base" (e.g. Chou, et al., 2004).

3. MCDM PROCEDURE FOR TECHNOLOGY SELECTION

In this section, we outline the basic steps used in MCDM as they apply to the technology selection. It consists of the following steps:

1. Definition of the problem and its alternative solutions
2. Identification of the stakeholders
3. Definition of selection criteria
4. Selection of the technique of preferences aggregation
5. Evaluation of solutions in respect to each selection criterion
6. Search for consensual solution

To illustrate the MCDM approach, we use the mobile payment industry as the running example. In the light of the many past mobile payment system failures, we consider this market to be an interesting case study for our illustration. Moreover, we assume that there is a real need to analyze and compare the different technologies using a more structured approach. There are also various important actors to include in our model. The propose analysis is just for illustration purposes. The data that are used for this analysis are derived from an extensive research in the literature, from opinions of few experts, and also from interviews conducted in Switzerland for the purpose of a previous published research. This was done as first data inputs to run the model in order to evaluate the pertinence of MCDM method for assessing the mobile payment market.

Definition of the problem and its alternative solutions. With the continuous growth of the mobile industry, a variety of wireless technologies emerged in order to enable new mobile products and services. Some of them

could be well used for mobile payment services. However, these technologies differ not only in their capabilities but also in the impact they have on the different stakeholders. For the purpose of this paper, we selected three technologies that are known to be leading candidates for mobile payments. The first alternative is the contactless card embedded with a RFID (Radio Frequency Identification) tag. These cards tend to become very popular for many reasons. They are cheap, reliable, and very easy to use. Then, we chose mobile phones using proximity networks such as Bluetooth, RFID, and Infrared. This solution is good for proximity payment in the real world. Finally, we included mobile phone using remote networks (e.g. GSM, GPRS, UTMS, EDGE, WLAN). These devices are suitable for remote payments such as e/m-commerce transactions. Each of the introduced technologies has its advantages and drawbacks. Some have limitations that others do not have. For that reason, a mobile payment service provider should consider all these options before launching a scheme. For benchmarking purposes, we also included two very popular payment technologies such as magnetic cards (e.g. VISA, Mastercard) and smartcards (e.g. Proton). This will help us not only to compare the new mobile with the existing technologies but also give us good insights about the current market.

Table 1. The Technologies Selected for the Mobile Payment Case

Technology	Example
Conctactless card (RFID tag)	Octopus (Hong Kong)
Mobile phone "proximity" (Bluetooth, RFID, Infrared)	Moneta (South Korea)
Mobile phone "remote" (GSM, GPRS, etc)	Paybox (Austria)
Magnetic card	Visa, Mastercard (Worldwide)
Smartcard	Proton (Belgium)

Identification of the stakeholders. As there are various identified stakeholders in the mobile payment industry, we classified them in different groups. We distinguish actors involved in mobile payment transactions directly (players) and indirectly (rulers). The rulers set a legal framework (regulators) by making rules and controlling others' obedience, while there are also diverse actors (technology suppliers) in charge of providing the technology to the players. On their side, the players represent the demand (merchants and consumers) and the supply (mobile payment service providers). For our research, we choose to include only the players as they are very important since the success of a mobile payment scheme necessarily depends on their adoption. Moreover, they must be convinced about a technological consensus in order to enhance the success of a particular solution.

For our analysis, we formed three stakeholder groups: *provider, merchant, and consumer*. We assume that there is a general consensus about the issues they individually worry about. Indeed, all the merchants agree

with each other. They are all part of the same group. The same assumption has been made with the other stakeholders.

Definition of selection criteria. Criteria are used to capture the points of view that decision-makers use as a frame for reference in their selection process. Criteria should be comprehensive in that when taken all together, they should be able to represent a rather complete perspective of the user with regard to the problem. Table 2 summarizes the criteria adopted by each of the three groups of stakeholders in the selection of mobile payment technologies. Due to space constraints, we just provide the list of the stakeholders criteria used for our case. They are derived from the literature.

Table 2. List of the stakeholders' criteria

Provider	Merchant	Consumers
Cost	Cost	Cost
Organizational change	Customer base	Ease of use
Security	Ease of use	Expressiveness
Standard	Reliability	Trust
	Security	Universality
	Value proposition improvement	Usefulness

Selection of the technique of preferences aggregation. As discussed earlier, MCDM allow analysis of several criteria simultaneously or concurrently. These criteria may be either quantifiable (e.g. cost, speed, etc) or non-quantifiable (e.g. quality of service, esthetics, etc). More importantly, at least from the decision-maker viewpoint, the multiple objectives often work against each other. The improvement or achievement of one criterion can be accomplished only at the expense of other.

MCDM also allow consideration of the decision-maker subjective evaluation which is often crucial in decision problems. In most MCDM, the decision-maker can express his/her preferences by weighting the evaluation criteria, making pairwise judgements or by simply giving an ordinal ranking of a subset of alternatives. The preference aggregation process can be algorithmically precise (e.g. multi-objective linear programming) or heuristic (e.g. spacial proximity). Quantitative or qualitative techniques such as simulation or scenario analysis can also be used as preliminary analysis prior to the use of a MCDM (Bui, 1987). Stewart (1992) and Salo et al. (2003) offer a review of some of the most popular MCDM techniques.

As an example of a technique of preferences aggregation, we chose ELECTRE I (Benayoun et al., 1966) for our mobile payment case. This approach allows the decision maker to select the ideal technology with a maximum of advantages and a minimum of inconveniences in the function of various criteria. ELECTRE I gives the possibility to model a decision making process by using the concordance and discordance indexes and the

outranking relations. The concordance index measures the degree of dominance of one action over another, based on the relative importance weightings of the decision criteria. The discordance index measures the degree to which an action is worse than another. In summary, concordance and discordance indices can be viewed as measurements of satisfaction and dissatisfaction that a decision maker senses when choosing one action over another.

The outranking relations are usually obtained with a combination of a high level of concordance and a low level of discordance. These levels are fixed by a concordance and a discordance threshold which can be seen as severity levels over and under which an action could outrank another.

Based on relatively simple hypotheses, the objective of ELECTRE I is modest as it simply proposes a subset of alternatives (in our case, technologies) which definitely excludes the "best" solution. As a result, the decision maker has to be conscious that the kernel (i.e. the set of non-dominated alternatives) includes not only the "best" solution but also all the alternatives that are hard to compare between each other.

Evaluation of solutions in respect to each selection criterion. The purpose here is to help the decision-maker express his preferences with regard to the possible solutions. This preference elicitation is made in respect to each of the criteria considered for the selection problem.

Table 3. Evaluation by the provider group (for illustration purpose)

Criteria	Weight	Magnetic card	Smartcard	Contactless Card	Mobile phone "remote"	Mobile phone "proximity"
Cost	60%	4	3	3	1	2
Org. chan.	10%	3	3	3	1	1
Security	10%	1	3	3	4	2
Standard	20%	4	3	1	2	1
0 = weak; 1 = fair; 2 = average; 3 = good; 4 = excellent						

Search for consensual solution. As previously claimed, a consensus between the major stakeholders of the market is desirable. As a result, the success probability of a global payment scheme based on a unanimous technology choice would be higher. Following this requirement, we had to use a group decision approach. Bui and Jarke (1984) have previously proposed a method based on ELECTRE I for group decision making. They suggested applying the min-max concept of game theory (von Neumann, 1953). In other words, to reach a consensus, this method takes the most severe technology evaluations for each criterion done by any actor.

Concordance Matrix

	Magnetic Card	Smartcard	Contactless Card	Mobile phone "Remote"	Mobile phone "Proximity"
Magnetic card	-	65	35	65	65
Smartcard	20	-	40	**85**	**85**
Contactless card	20	60	-	70	**100**
Mobile phone "Remote"	10	10	25	-	40
Mobile phone "Proximity"	10	0	20	65	-

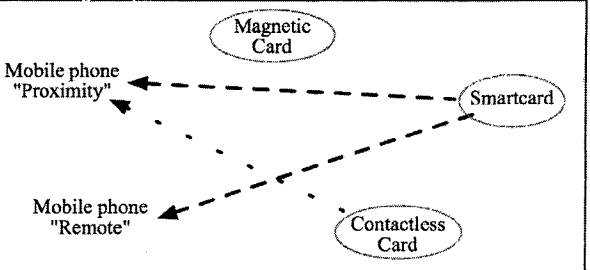
Discordance Matrix

	Magnetic Card	Smartcard	Contactless Card	Mobile phone "Remote"	Mobile phone "Proximity"
Magnetic card	-	56	63	67	44
Smartcard	**22**	-	33	**13**	**13**
Contactless card	70	60	-	40	**0**
Mobile phone "Remote"	78	56	63	-	33
Mobile phone "Proximity"	70	60	50	56	-

Outranking Matrix

(P=.75 and Q=.25)	Magnetic Card	Smartcard	Contactless Card	Mobile phone "Remote"	Mobile phone "Proximity"
Magnetic card	-	0	0	0	0
Smartcard	0	-	0	**1**	**1**
Contactless card	0	0	-	0	**1**
Mobile phone "Remote"	0	0	0	-	0
Mobile phone "Proximity"	0	0	0	0	-

Outranking Graph



Legend

Concordance matrix
100 = All criteria favor the selection
0 = No criteria favor the selection

Discordance matrix
0 = No opposition to selection
100 = Total opposition to selection

Outranking matrix
0 = No outranking relation
1 = Outranking relation

Outranking graph
A directed arrow shows that a solution outranks the other

Figure 1. Group result (illustration)

4. SUMMARY

The use of MCDM for assessing the mobile payment market shows encouraging results. In fact, even with our first exploratory data inputs, we obtained interesting insights that are somewhat representative of the current market state.

The proposed methodology should improve the technology selection process as it allows consideration for multiple actors to express their

personal preferences in the selection process. With the use of a DSS, we could perform sensitivity analysis and visualization of the outcome capabilities. DSS also present the possibility to conduct market simulations in order to build evolving scenarios.

In conclusion, we hope that using MCDM could be a key component for improving the development of successful mobile information systems, as the choice of technology is crucial.

A further research would be to apply this approach in a real setting and therefore, capture the stakeholders' preferences of the current market.

ACKNOWLEDGEMENTS

The work presented in this paper was supported by the National Competence Center in Research on Mobile Information and Communication Systems (NCCR MICS), a center supported by the Swiss National Science Foundation under grant number 5005-67322.

References

Benayoun, B. , Roy, B., and Sussmann, B. (1966). Manuel de reference du programme electre. *Note de synthèse, formation n.25, Direction scientifique SEMA, Paris.*

Bui, T. (1987). *Co-oP: A Group Decision Support System for Cooperative Multiple Criteria Group Decision Making*, volume 290 of Lecture Notes in Computer Science. Springer-Verlag, Berlin.

Bui, T., Jarke, M. (1984). A dss for cooperative multiple criteria group decision making. *International Conference on Information Systems (ICIS)*

Chou, Y., Lee, C., and Chung, J. (2004). Understanding m-commerce payment systems through the analytic hierarchy process. *Journal of Business Research*, 57, 12, 1423–1430.

Chan, F.T.S, Chan, M.H., and Tang, N.K.H. (2000). Evaluation methodologies for technology selection. *Journal of Materials Processing Technology*, 107, 330-337.

Keen, P.G.W. (1977). The involving concept of optimality. *TIMS Studies in the Management Sciences*, 6, 31-57.

Von Neumann J. and Morgenstern O. (1953). *Theory of games and economic behavior*, 3rd Ed. Princeton University Press, Princeton, New Jersey.

Saaty, T. (1980). *The analytic hierarchy process: planning, priority, allocation.* McGraw-Hill Book Company, New York.

Salo, A., Gustafsson, T., and Ramanathan, R. (2003). Multicriteria methods for technology foresight. *Journal of Forecasting*, 22(2):235–255.

Stewart, TJ. (1992). A critical survey of the status of multiple criteria decision making theory and practice. OMEGA, 20, 569–586.

Zeleny, M. (1982). *Multiple objective decision making.* Addison-Wesley, Readying, Massachusetts.

MOBILE SYSTEMS DEVELOPMENT

Challenges, Implications and Issues

Jens Henrik Hosbond
Aalborg University, Dept. of Computer Science

Abstract: This paper takes a systems development perspective on mobility, building on preliminary findings of an on-going multiple case study covering 7 companies. The questions driving this paper are: What are the challenges facing development practice in the mobile industry, how do they affect practice and how are they dealt with? Analysis of the empirical data is done following a structured and inductive approach. A framework showing the segmentation of the mobile industry into five layers is proposed and challenges are presented according to two dimensions, namely a business dimension and a development dimension. Finally, implications stemming from these challenges are discussed and issues inviting for future research are proposed.

Keywords: Mobile systems development, multiple case study, software development

1. INTRODUCTION

Development of mobile systems is a challenging task surrounded by a high level of uncertainty: Rapid technology development and lacking standardisation, short time-to-market, lacking end-user adoption, missing killer applications are just some of the conditions creating an uncertain environment. Despite, the much uncertainty surrounding mobile systems development (MSD), contributions on the subject have been sparse - see e.g. (Lyytinen, Rose et al. 1998; Krogstie 2001; Krogstie, Lyytinen et al. 2004; Hosbond and Nielsen 2005). Obviously, traditional systems development (TSD) (Wieringa 1998; Sommerville 2000; Presssman 2004) should have a role to play in trying to understand MSD, but the limitations to its applicability are unclear. This study seeks to uncover the challenges that surround MSD practice and the actions taken in practice to handle these.

Respectively, it must be mentioned that Krogstie et al. (2004) have added to a greater understanding of the challenges in MSD on a conceptual level even though it remains unclear to what extent their propositions are based on empirical data. Not before understanding the challenges of MSD practice are we able to reason about the limitations of TSD and suggest improvements. Thus, the questions driving this paper are: What are the challenges facing development practice in the mobile industry, how do they affect practice and how are they dealt with?

The next section outlines the research approach. Section 3 is a presentation and analysis of the cases. Section 4 is a discussion of the analysis and its implications and Section 5 concludes the paper.

2. RESEARCH APPROACH

MSD is indeed still uncharted land. To reach a holistic level of understanding we set out to (1) explore MSD challenges on different segments of the mobile industry and (2) to understand the dynamic interrelations between businesses and its effects on development practice. Hence, an exploratory multiple case study approach is applied (Yin 1994). Unit of analysis is development practice in seven companies within the Danish mobile industry. Collection of qualitative data is conducted using semi-structured and open-ended interviews (Patton 1990; Yin 1994). Questions address company background, development activities, and external collaborative efforts. Interviews are audio recorded and transcribed for later analysis.

Analysis of the empirical data is done following a structured and inductive approach taking place in between interviews, allowing for an "overlapping data analysis [and having] the freedom to make adjustments during the data collection process" (Eisenhardt 1989, pp. 539). To generate explanatory theory from case studies a disciplined and structured approach to data analysis must be followed (Eisenhardt 1989; Dyer and Wilkins 1991). Effectively, a structured coding practice is initiated, applying the software coding tool Hyperresearch supporting the practice of inductive analysis. Through coding of the transcripts concepts are added as a level of abstraction to represent challenges. The added concepts emerging from the inductive analysis are used as constituting elements of the explanatory theory that this paper seeks to build (Eisenhardt 1989).

3. CASES AND ANALYSIS

Early on in the data collection and data analysis process it becomes clear that the mobile industry is divided into a set of layers forming a supply-chain. These constituting layers (see Figure 1) and their interrelations are important to understand as they may affect and indirectly shape development practice in other layers.

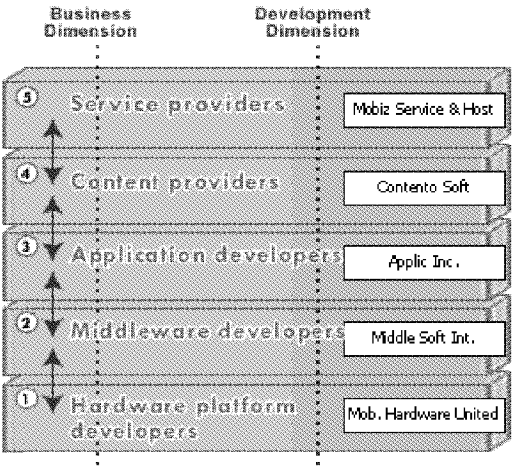


Figure 1. The five layers constituting the mobile industry.

Through analysis of the data, two dimensions emerge; the business dimension (business related challenges) and the development dimension (development specific challenges), see Figure 1. Together they represent the analytical lens applied for presenting and discussing the challenges, see Table 1. Cases are given fictive names for reasons of confidentiality.

Table 1. Overview of cases, business- and development related challenges.

Description	Business dimension	Development dimension
▪ Layer / Case: 1 / Mobile Hardware United ▪ Profile: One out of six R&D sites within the company. Employing 150 hardware- and software engineers. A dominant global player within this segment of the mobile industry. ▪ Products: High quality hardware platforms and software for testing reliability and correctness of hardware platforms. ▪ Supply-chain: Dependent on other R&D sites within the company as development takes	▪ Tough competition on price	▪ Communication and alignment of requirements across development projects ▪ Cultural differences in interorg. development. ▪ Quality management

Description	Business dimension	Development dimension
place through umbrella projects. ▪ Interviewee: Software director responsible for initiation and completion of development projects part of interorg. development projects.		
▪ Layer / Case: 2 / Middle Soft International ▪ Profile: One out of 9 R&D sites around the world. Employing 400 software engineers. Global player involved in all layers except for layer 5. ▪ Products: Software platform and client application development. ▪ Supply-chain: Dependent on software development projects at other R&D sites within the company. ▪ Interviewee: Line manager, responsible for improving software development processes and alignment and coordination of development initiatives across R&D sites.	▪ Conflicting customer interests (mobile operators vs. end-users): revenue creating vs. free communication enabling technologies ▪ Competition: Time-to-market and price	▪ Dispersed development: Alignment of requirements and development process ▪ Quality management and security ▪ Rapidly changing technology
▪ Layer / Case: 3 / Applic Inc. ▪ Profile: Small company originating from Nokia Denmark R&D. Customers are from layer one, two, and three. ▪ Products: Consultancy business, in-sourcing of project managers and systems developers, out-sourcing of entire development projects, embedded software development. ▪ Supply-chain: Dependent on establishing and maintaining good business relations with customers and strategic partners. ▪ Interviewee: Technical director responsible for initiation and completion of development projects.	▪ Competition: Time-to-market and price	▪ Lacking standardisation and documentation of middleware. ▪ Integration and development on top of proprietary software platforms (middleware).
▪ Layer / Case: 4 / Contento Soft ▪ Profile: Telecommunications branch within a large Danish Software development company employing around 2500 people. ▪ Products: Managing and delivering sensitive content to mobile services targeted at jobs in public institutions, e.g. for home care workers. ▪ Supply-chain: Dependent on formulation of revenue share and development cost share	▪ Revenue- and especially development cost-share models with mobile operators to lower business risks. ▪ Lacking adoption of	▪ Lacking standardisation of proprietary middleware and mobile applications. ▪ Integration with existing "static" IT systems. ▪ Data security

Description	Business dimension	Development dimension
business models. ▪ Interviewee: CIO in the telecommunications branch of the company.	mobile services.	▪ Defensive development strategy due to rapid technology replacement.
▪ Layer / Case: 5 / Mobiz Service & Host ▪ Profile: Employs 60 people in Denmark, Germany, and the UK. "Buys" applications (layer 3) and content (layer 4) and "sells" to mobile network operators (layer 5). Customers are mobile operators. ▪ Products: Offers mobile applications and hosting and service management of these. ▪ Supply-chain: Highly dependent on good business relations with application developers and content providers (layer 3 and 4) and mobile network operators on layer 5. ▪ Interviewee: Technical project manager responsible for integration and adaptation of mobile services onto software platforms and infrastructure components.	▪ Establishment of good revenue-share models is critical. ▪ Lacking adoption of mobile services making service and hosting of mobile services less sellable.	▪ Lacking standardisation of proprietary software platforms. Tailored solutions are necessary. ▪ Adaptation and integration of software components instead of bottom-up development.

4. DISCUSSION

The challenges from the analysis together with derived implications and issues are summarised in Table 2. Business related challenges are included in Table 2 as these affect the conditions under which development projects take place. Discussion of these challenges and their implications infer issues needed to be addressed in order to create further understanding of these challenges and implications. Implications and issues stemming from interorganisational development and lacking standardization are now further elaborated on.

Table 2. Challenges, implications and issues.

Challenges	Implications	Issues
▪ Software quality ▪ Competing on price ▪ Software innovation ▪ Time-to-market ▪ Changing technology	▪ Highly complex development practice: Efficient, structured, agile, and enabling innovation.	▪ Mixing structured and process-oriented with agile and product-oriented development methods.

Challenges	Implications	Issues
▪ Interorganisational development	▪ Development projects exceed organizational boundaries.	▪ Extension of project scope in systems development. ▪ Defining, coordinating and aligning requirements. ▪ Cultural differences.
▪ Lacking standardisation	▪ Enormous resources spent on adaptation and integration.	▪ Contemporary development methods in TSD assume bottom-up approaches – not integration and adaptation.

4.1 Interorganizational Development and Lacking Standardisation

Development practice in the mobile industry is characterised by dispersed R&D sites in multi-national corporations such as Mobile Hardware (layer 1) and Middle Soft Int. (layer 2). Development projects do not merely take place within the physical boundaries of a site. Instead development projects are defined as umbrella projects each comprising an arbitrary number of sub-projects with different responsibilities and tasks. This phenomenon is mostly related to businesses residing in the "mature" segments of the mobile industry. In the upper layers of the mobile industry (layer 3 to 4), collaborative partnerships and strategic alliances around development activities tends to dominate. The most obvious reason for this tendency is the missing critical mass of users that has not yet come to adopt and accept paying mobile services (Ling 2000; Aarnio, Enkenberg et al. 2002; Naruse 2003). Financial resources are therefore sparse and with a vast number of wireless technologies, businesses tend to specialise on a few core services and product areas. To strengthen the ability to invent smarter and better-looking services and content, businesses are to a greater extent establishing collaborative partnerships and strategic alliances around product development to stay competitive. Effectively, this means that development projects no longer can be perceived as physically fixed and taking place within a predefined team structure. Development activities exceed organisational boundaries as tasks are divided between the involved development teams. From interorganisational development, issues such as aligning requirements and solving cultural differences emerge. However, taking a look at the contemporary literature on traditional systems development, e.g. (Wieringa 1998; Sommerville 2000; Presssman 2004), it may be argued that the described development methods implicitly assume development teams to be geographically fixed and a development practice limited to the boundaries of single organisation. Consequently, a critical

stance as to the cultural issues is not provided. To enhance the understanding of interorganisational development we suggest an extension of the development project scope in mobile systems development taking an interorganisational perspective on MSD (Hosbond and Nielsen 2005). Furthermore, research within organisation of open source projects may prove beneficial to understand the phenomenon of interorganisational development. What are the mechanisms and structures applied in coordinating and aligning open source projects and can these mechanisms be applied in the attempt to understand the issue of interorganisational development in greater detail?

The mobile industry is still largely technology-driven. Focus is on innovation instead of standardisation. This is manifested in the vast amount of competing wireless technologies enabling mobile work, communication, collaboration etc. Technologies are innovated at the lower layers in the industry (layer 1 and 2) but it is the upper layers that are struggling with the consequences hereof. According to a report by Forrester Research application developers and content providers are forced to target their development against several SW platforms such as Java, Symbian, Windows Mobile, and Brew (Lussanet, Østergaard et al. 2004) in order to reach as many end-users as possible. In addition, middleware developers such as Middle Soft International extend the raw Symbian platform with proprietary features – despite that one of the core ideas of Symbian is to establish a standardised platform. The strong focus on innovation instead of standardisation at the lower layers of the industry implies that enormous resources are spent on integration and adaptation of applications and content on top of the SW platform. From a systems development point of view we may pose the question of how to actually approach integration and adaptation in systems development. It seems that traditional systems development literature does not lend it self to discussing these matters. Instead a bottom-up perspective on systems analysis, design, and implementation is often assumed.

5. CONCLUSION

This paper presents preliminary findings of an on-going multiple case study. Based on an inductive analysis applying a structured coding scheme using the analysis tool Hyperresearch, a five layered framework of the mobile industry is proposed and challenges in development practice are elaborated on. The challenges are discussed with respect to the implications for development practice and several issues, see Table 2, inviting for future research for mobile systems development are suggested.

References

Dyer, W. G. J. and A. L. Wilkins (1991). "Better Stories, Not Better Constructs, To Generate Better Theory: A Rejoinder To Eisenhardt." Academy of Management Review **16**(3): 613-619.

Eisenhardt, K. M. (1989). "Building Theories from Case Study Research." The Academy of Management Review **14**(4): 532-550.

Hosbond, J. H. and P. A. Nielsen (2005). Mobile Systems Development - A literature review. Proceedings of IFIP 8.2 Annual Conference, Cleveland, Ohio, USA, IEEE.

Krogstie, J. (2001). Requirements Engineering for Mobile Information Systems. Proceedings of REFSQ'2001, Interlaken, Switzerland.

Krogstie, J., K. Lyytinen, et al. (2004). "Research areas and challenges for mobile information systems." International Journal of Mobile Communications **2**(3): 220-234.

Ling, R. (2000). ""We will be reached": the use of mobile telephony among Norwegian youth." Information Technology and People **13**(2): 102-120.

Lussanet, d. M., B. Østergaard, et al. (2004). Mobilizing Content For 3G Delivery, Forrester Research: pp. 18.

Lyytinen, K., G. Rose, et al. (1998). "The brave new world of development in internetwork computing architecture (InterNCA): or how distributed computing platforms will change systems development." Information Systems Journal **8**: 241-253.

Naruse, K. (2003). The survey of the mobile Internet, usage, awareness, study for m-commerce. Proceedings of the 2003 Symposium on Applications and the Internet Workshops (SAINT'03 Workshops), Orlando, Florida, IEEE.

Patton, M. Q. (1990). Qualitative Evaluation and Research Methods, SAGE Publications.

Presssman, R. S. (2004). Software engineering - a practitioners approach. New York, McGraw-Hill.

Sommerville, I. (2000). Software Engineering. Harlow, England, Addison-Wesley.

Wieringa, R. (1998). "A Survey of Structured and Object-Oriented Software Specification Methods and Techniques." ACM Computing Surveys **30**(4): 459-527.

Yin, R. K. (1994). Case study research - design and methods. Thousand Oaks, California, SAGE Publications.

Aarnio, A., A. Enkenberg, et al. (2002). Adoption and Use of Mobile Services - Empirical Evidence from a Finnish Survey. Proceedings of the 35th Annual Hawaii International Conference on System Sciences, Big Island, Hawaii, IEEE.

A SYSTEM FOR MOBILE AND WIRELESS ADVERTISING

Personalized information as incentive for receiving advertisement on mobile terminals

Michael Decker[1], Rebecca Bulander[1], Gunther Schiefer[1] and Bernhard Kölmel[2]

[1]*Institute AIFB, University of Karlsruhe (TH), 76 128 Karlsruhe, Germany;* [2]*CAS Software AG, Wilhelm-Schickard-Str. 10-12, 76 131 Karlsruhe, Germany*

Abstract: Mobile terminals are an interesting medium for advertising because of the high penetration rates and their character as personal communication devices. Since advertising in general has the reputation of being something annoying there has to be some kind of incentive mechanism to obtain permission from the consumers for advertising on their mobile terminals. The MoMa-system for mobile and wireless advertising described in the following article focuses on personalized information as such an incentive mechanism. For the provision of personalized information a system requires personal data like profile information and the current location of an user. But there are privacy concerns when providing such information for a mobile advertising application, of course. Thus we designed MoMa in a way to realize both of this conflicting requirement: personalized information and guaranteeing data protection.

Keywords: Mobile and wireless advertising, context sensitive mobile applications, data protection

1. INTRODUCTION

Advertising is defined as making a representation in any form in connection with trade, business, craft or profession in order to promote the supply of goods or services, including immovable property, rights and obligations (directive 84/450 of the European Economic Community). Kotler and Bliemel's (1992) definition of advertising comprehends a paid and non-

personal representation. Based on these definitions mobile or wireless advertising is advertising using mobile terminals as medium. While some authors consider on-board vehicle computers or notebooks as mobile terminals, we restrict ourselves to mobile handheld devices like cellular phones, PDAs and smartphones. Notebooks and on-board computers don't have the ubiquitous character of handheld devices: one doesn't carry his notebook with him all day, and even if one does, it won't be turned on all the time (Turowski and Pousttchi, 2004).

Mobile advertising can be seen as advancement of digital- or internet-advertising, since both have special features in common which are valuable for advertising purposes; but the ubiquitous character of mobile terminals even increases the potential of these features.

Although the term "mobile advertising" (e.g. Barwise and Strong, 2002) as well as "wireless advertising" (e.g. Yunos et al, 2003; Wohlfahrt, 2001; or the Wireless Advertising Association) can be found in literature, strictly speaking one should use the term "mobile *and* wireless advertising" (figure 1), since "mobile" and "wireless" are orthogonal concepts (Wang, 2003): if a device is mobile (one can easily move it) or not doesn't imply if its data connection is wireless or wired. An ordinary desktop PC (not mobile) might be connected to the internet via Ethernet (wired) or using wireless technologies like WiFi or WiMAX because someone doesn't want to lay cables in the whole house respective there is no cable for broadband internet access available in his street. A notebook or a simple PDA without WiFi or Bluetooth capabilities are examples for wired mobile devices. Services like the one by AvantGo (avantgo.com) show that these kinds of devices can also be used as medium for advertising: using AvantGo owners of PDAs can download web pages from the ordinary internet for offline reading and sync them to their devices; the pages are transcoded to be readable on small PDAs and adverts are added. For the sake of simplicity we will use the term "mobile advertising" when we actually mean "mobile and wireless advertising".

The rest of this article is organized as follows: in the second chapter we discuss selected aspects of mobile advertising. Chapter number three describes the MoMa-system, which was developed during the project "MoMa — Mobile Marketing" funded by the Federal Ministry of Economics and Labour of Germany. The last chapter gives a summary.

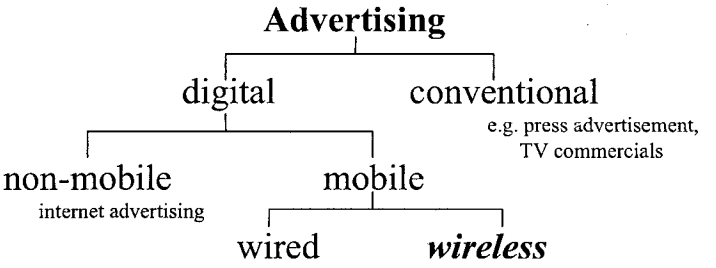


Figure 1. Different methods of advertising

2. MOBILE ADVERTISING

2.1 Potentials of Mobile Advertising

Mobile terminals are an interesting target platform for advertising because of the high penetration rates: in many countries the penetration rate exceeds the 90 % mark, there are even countries like Sweden and Italy with mobile penetration rates over 100 % (Netsize, 2005). The global number of mobile subscriptions is expected to pass the number of two billions during 2005, and approaching 3 billions by the end of the decade (Informa Telecoms & Media, 2005). Also they are no longer considered as extraordinary gadgets but as commonplace household commodities.

People carry their mobile device with them most of the day, seldom lend it away and don't share it with other people (which is quite common for other communication devices, e.g. one telephone or TV set for the whole family), so using mobile technologies advertisers can reach people almost anytime and anywhere. During the pretest for an acceptance study for a mobile service we asked the participants (N=28) how many hours a day on average they are reachable via their cellular phone, and the mean value of the answers was ≈ 20.7 h/day (deviation ≈ 5.1), whereas ≈ 56 % stated "24 h". Using conventional advertising methods a marketer can reach his audience only in certain situations and moments, e.g. after work in front of the TV set when using commercials.

Mobile advertising can be seen as advancement of internet advertising, whereas the market volume in 2004 for the latter is estimated to be as high as 9.6 billion US $ alone for the US-market (IAB, 2005). Both share the

following features, but the ubiquitous and personal character of mobile devices even increases the advertising potentials provided by these features:

- Individually addressable: Mobile terminals can be addressed individually (using the mobile phone number), so personalized and target-oriented advertising can be realized. This is also the case for internet advertising (using the e-Mail- or IP-address). Using conventional advertising methods the advertiser has to broadcast his message to an anonymous crowd of people (e.g. audience of a TV commercial or readership of a newspaper or magazine) so many people are reached that are not in the intended target audience. Advertising messages on mobile terminals are hard to ignore when using push-mode communication (e.g. SMS/MMS or text-to-speech).
- Interactive: Internet and mobile advertising can be interactive. One can request further information, participate in a sweepstake or forward the message to friends without changing the medium (e.g. when sent as e-mail or SMS). Especially "forward a message to friends" is an interesting opportunity from a marketing point of view; there are even campaigns with the aim to stimulate a "viral"- or "word-of-mouth"-effect (e.g. "Wella Virtual Kiss", Mohr et al, 2003): an advertisement message is passed to a consumer by another consumer, not by the advertiser directly (Helm, 2000). Viral marketing is supposed to provide a dissemination of adverts with exponential growth and it is assumed that consumers have more confidence in advertisement messages received from friends rather than firms; there are special systems for mobile advertising following the idea of "viral marketing" (Straub and Heinemann, 2004; Ratsimor et al., 2003) using multihop ad hoc networks without infrastructure (so called MANETs, see Toh (2002)). Also the response rates reported for mobile advertising are very promising (Kavassalis et al., 2002; Schwarz, 2001), many users respond within a few minutes after they received an advert (Barwise and Strong, 2002).
- Multimedia capabilities: internet advertising may make use of multimedia content like pictures, jingles, sounds, animated graphics and movie sequences. This is very important for advertising since many marketers use such elements to create brand awareness. Many mobile terminals used today already can display multimedia content.

A special feature of mobile services is context awareness: Context sensitive applications use information about the situation of an user to adapt themselves according to his needs in that situation (Schilit et al., 1994; Chen and Kotz, 2000). The most prominent example for context information concerning mobile applications is "location". Technically the information about the current location of an end user respective his mobile device can be

retrieved using a receiver for the signal of the global positioning system (GPS), referring to the location of the nearest base station (cell-of-origin) or combining both methods to obtain "assisted GPS". The "time difference of arrival"-method (TDOA) calculates the current position referring to the observed time difference of arrival of radio signals from several base stations (Zeimpekis et al., 2003). Applying the idea of location based services (LBS) to advertising one could provide an user with advertisement concerning facilities not far away from his current location (Aalto et al., 2004; Kölmel and Alexakis, 2002) or even suggest a route to get there (Ververidis and Polyzos, 2002). But "location" is just one example for context information; there are many other examples like time or weather and MoMa is designed to handle all kinds of thinkable context parameters. The usage of context information is very important when designing a mobile application, because mobile terminals have a limited user interface (no real keyboard, small screen) and thus it is import to decrease the amount of information that has to be entered by the user to use a mobile service.

The "honeymoon effect" — people are interested in new (technical) things — could also help to establish mobile advertising.

2.2 Challenges of Mobile Advertising

In the last subchapter we discussed the potentials or advantages of mobile terminals as advertising platform. But there are also big challenges when talking about mobile advertising:

Since mobile advertising can be seen as advancement of internet advertising there is the concern that the huge wave of Spam messages known from e-mail-communication spills over to mobile devices. Spam is defined as an unsolicited electronic bulk message with commercial intent (OECD, 2004). There are studies that state values of more than 70 % for the portion of spam message in e-mail-communication, for example MessageLabs (2004). Another study by bmd wireless and the University of St. Gallen (2005) found out that 80 % of the people questioned already received Spam messages on their cellular phones. Besides spam messages there are other kinds of unsolicited messages: messages which pretend to be personal messages and ask the user to call back to a certain number (which is in fact premium rate number), contain a virus or which change the configuration of the handset.

Unsolicited messages for mobile terminals are an even bigger problem than for desktop PCs because of the limited resources and the personal or even intimate nature of mobile devices; most mobile devices haven't enough memory to store a lot of unsolicited messages or don't have the computation power necessary for running a spam filter. As our mentioned pretest-study

implies there are a lot of people who never turn of their cellular phone, so an unsolicited message could even disturb their sleep.

There are also some special challenges for mobile advertising due to the nature of mobile terminals:

- Limited user interface: Due to their limited size and weight mobile terminals have a limited user interface: they have a small display with a low resolution and color depth and don't have a full keyboard, so users don't want to enter a lot of data and advertising messages have to be designed in a way to be displayable in a reasonable way on mobile terminals. We consider context information as discussed above as a way to relieve the user of entering more data than necessary.
- Limited resources: mobile terminals don't have big resources with regard to bandwidth, calculation power, memory and — probably the worst problem — battery power.
- Expenses of mobile data communication: mobile data communication is still very expensive, e.g. about one Euro for 1 Mbyte data traffic (cleared in blocks of 10 or even 100 KByte) when using GPRS or UMTS (prepaid rates are even much more expensive), so many people don't use mobile devices for research on product and services. Also nobody wants to pay for the reception of advertisement.
- Privacy concerns: People store sensitive data on their mobile devices (e.g. address book, calendar, personal notes) and it is possible to track their location (see the discussion for location based services in section 2.1), so it's no wonder that there are privacy concerns, e.g. Barkhuss and Dey (2003). There are also laws which ask for data protection, e.g. directive 95/46/EC of the European Economic Community.
- Different types of mobile terminals: There is a plethora of different types of mobile terminals on the market, all with different capabilities with regard to color or monochrome display, display size and resolution, devices for data input; see the WURFL (2005) project, which maintains an open database with specifications of many different mobile devices. An advert that looks great on one type of device may look terrible on another one (if presentable at all). When developing a mobile application there are high costs caused for testing and porting the application to different types of terminals; these costs often exceed the costs for the actual implementation (Schlickum, 2005).

3. THE MOMA SYSTEM

3.1 Basic Principle

A fundamental concept in mobile advertising due to the experience with unsolicited direct advertising — in particular spam-e-Mail and telephone calls ("cold calls") — is permission marketing (Godin, 1999; Krishnamurthy, 2001): consumers will only receive ads after they have explicitly expressed their allowance; they can opt-out anytime if they no longer want to receive advertisement, of course. In many countries permission marketing is the only legal way of direct advertising (e.g. article 13 of the directive 2002/57/EC of the European Union). Permission marketing also asks for personalized ads which fit the fields of interest (profile) of the consumer.

There is one hitch with the concept of permission marketing: a consumer has to know about a brand or firm to give explicitly allowance for receiving their adverts. So firms have to employ conventional methods of advertising (e.g. TV commercials) to "invite" consumers for participation in a mobile advertising campaign (Kavassalis et al., 2003). For this reason mobile advertising is often integrated in bigger campaigns, see Bauer et al. (2005) for examples. However, small enterprises don't have the resources to do this.

To realize permission based mobile advertising also affordable for small enterprises we designed MoMa as mediator between advertisers and end users (see figure 2): on the right side of the system advertisers put "offers" into the system. These offers are formulated according to a "catalogue" which is a tree of product and services categories. Each category is specified by certain attributes and inherits attributes of its parent. For example "gastronomy" with the attribute "price level" has "pubs", "restaurants" and "catering services" as child categories. On the left side of the system the end users submit "orders" to MoMa using a special client application on his mobile device; there are client applications for different types of mobile programming platforms (J2ME, Symbian). These orders are also formulated according to the catalogue and may include profile and private context information.

The matching-component of MoMa tries to find fitting orders and offers; for this process public context information may be requested from special context providers. If matches are found the end users of the order is notified (but not the advertisers!). These notifications are delivered using the channel defined in the end user's notification profile; such a notification profile describes how (SMS, MMS, e-mail, text-to-speech) and to which end address he wants to be notified. Different channels and end addresses can be

used depending on the time, e.g. e-Mail during bedtime, SMS to phone number A from 8 to 9 a.m and to phone number B from 3 to 4 p.m.

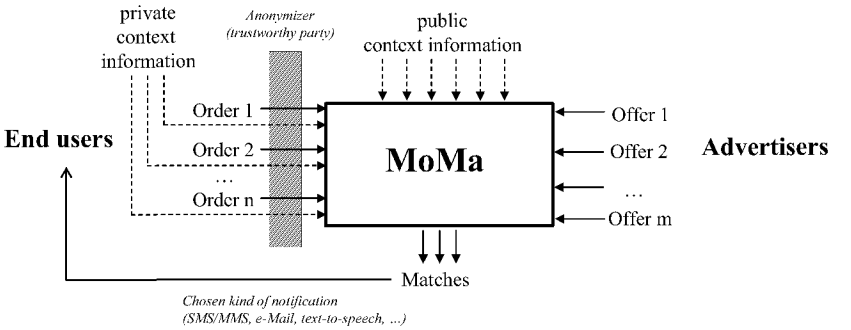


Figure 2. Basic principle of the MoMa-system

3.2 System Details

Each end user (consumer) of the MoMa-systems (see figure 3) needs an account with an unique user-ID and at least one general and one notification profile. In the general profile information about the user (age, marital status, ...) and his fields of interests are stored. A notification profile holds at least one end address of an user (e-mail-address, mobile phone number, ...); in the case of multiple end addresses the user can specify, what end address at what time interval has to be used for notification. Both profiles can be stored on a server and can be synchronized with different terminals an user may own.

When creating an order X the user has to choose a category from the catalogue and to fill in the needed attributes. If possible values for attributes will be looked up in the general profile (e.g. number of children when looking for hotel accommodation) or the available private context parameters (e.g. location). Please note: the order X does *not* contain information about the identity or the end addresses of an user.

The completed order X together with the user-ID (UID) and the number of the chosen notification profile (NID) are submitted to the anonymizer component of the trustworthy party. Appending a random bit string *rand* the anonymizer encrypts this and obtains a cipher text C=*crypt(UID, NID, rand)*. The random data included ensures that even when using the same ID and notification profile number multiple times we obtain a different cipher text. The pair {X,C} is then forwarded to the core component of the MoMa-system, which cannot decrypt C. This C can be seen as transaction

pseudonym, a pseudonym that is only used for one transaction and thus represents the most secure level of pseudonymity (Pfitzmann and Köhntopp, 2000).

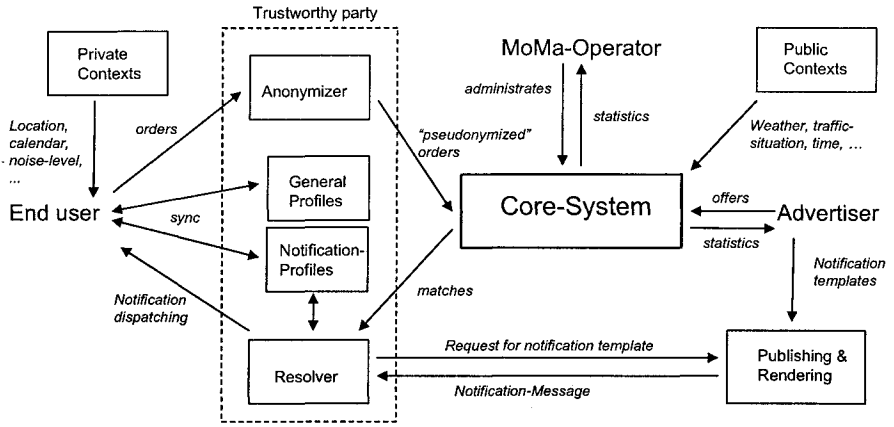


Figure 3. Architecture of the MoMa-system

On the other side the advertiser defines his offer Y according to the catalogue and submits it to the core component directly. He also has to create templates for the notification of end users concerning his offer Y and deposit them on the publishing & rendering server.

Within the core component of the MoMa system the matching server tries to find matching pairs of orders and offers. For each detected matching pair X and Y the core component sends {{X,C},Y} to the resolver component of the trustworthy party. C is decrypted to obtain the relevant notification profile (type of channel and corresponding end address). Afterwards the resolver requests the notification template from the publishing & rendering server, creates the notification message and dispatches it. The resolver component can add suitable labels to the notification messages to guarantee that they are always identifiable as advertising, which is important because of legal requirements.

There are several cases when it is necessary to change an order X: the user alters attributes of the order, the user wants to suspend or delete an order, or private context parameters have changed. In this case the updated X along with C (which is the same as for the original order) is sent to the MoMa-core-component, which looks up the old order X by its C and replaces respective deletes it.

3.3 Matching Component

For the implementation of the matching server it turned out to be helpful to employ the paradigm of "software agents" (Genesereth, 1994). An agent in this sense of computer science is a software component that resides in a special runtime environment (agent container), has a certain degree of autonomy and intelligence and is able to communicate with other agents by exchanging messages in a certain language.

Each order is represented by an instance of an order agent and may encapsulate rules, e.g. a gastronomy-order shouldn't match with a "beer garden" when the public context parameter "weather" concerning the current private context parameter "location" is "rainy". All instances of offers are represented by a single instance of an offer agent. For each type of public context information available there is one agent which can be queried by the other agents. In the case of certain events (updated orders/offers or context parameters, new orders and offers) a short-lived notification agent whose purpose is to communicate the change to the affected agents will be created. If an order agent detects an offer suitable for him he will contact the notification agents, who will initiate the notification process. Orders may be configured to have an expiry date, so order agents can kill themselves.

While there is the concept of mobile agents — agents can move themselves from one instance of the runtime environment or platform to another — for MoMa stationary agents were used. From an architectural point of view it would be preferable to have mobile agents and thus create the agents representing an order or offer on the mobile device respective a server of the advertiser and then move them to the matching platform; but this would require much more resources on the mobile terminals of the end users and each advertiser had to maintain his own agent platform, so we used the stationary agents for the implementation of MoMa.

3.4 Business Model

The basic business model for each form of advertising is that the advertisers have to pay for the presentation of their adverts. This is also the case for the MoMa-approach: the advertisers have to pay for each contact (notification message to an end user) generated by one of their offers. The price for one contact depends on the category of that order and will be in the magnitude of a few Euro-Cents for most categories. For categories covering very costly goods and services (e.g. real estates) or with a lot of competitors higher prices are thinkable. The current implementation assumes the same contact price for a given category for all advertisers; if the number of matching offers for an order is bigger than the number of notification

allowed by that end user we choose the orders to be displayed by random. Another approach would be to allow the advertisers to set a price or bid for a contact and to choose the offers with the highest contact prices.

An additional source of revenue for the MoMa-operator is selling statistical data concerning the pattern of demands, e.g. what kind of product where most often requested by the end users.

In contrast to many conventional method of advertising we do not have to estimate the number of contacts generated like for example for television commercials or newspapers adverts; the advertiser has only to pay for real contacts. There is no requirement to buy a big deal of contacts like for TV commercials, so MoMa-advertising is even affordable for individual enterprises. Also contacts generated by MoMa reach people that explicitly expressed their interest for a certain kind of product or service.

If the end users perceive the offers delivered by MoMa because of their highly personalized nature as valuable information rather than advertisement it is even thinkable that they are willing to pay for the MoMa-service. But in the current implementation they only have to pay the costs for the data transmission when putting an order into the system. Although the prototypes of the MoMa-client-applications use webservice technologies for the communication with the MoMa system and as the usage of a webservice causes a lot of data overhead the data volume for the transmission of one order is less than 1 KByte, so these costs are negligible or will be in the nearer future.

The providers for the public context information (e.g. specialized news agencies for weather or sport events) will be paid by the MoMa-operator. The trustworthy third party could be compensated by the MoMa-operator or could also be a government institution or a non-profit association.

When introducing a system like MoMa there is the well known "chicken-and-egg" of obtaining the necessary critical mass for end users as well as advertisers. End users will only use MoMa if there are enough offers in the system. But advertisers will only put offers into the system if there are enough end users of MoMa. To overcome this vicious circle there is the possibility of automatically obtaining offers from well established eCommerce-platforms without charging the operators of those platforms. Since many eCommerce-platforms offer special interfaces to do this (e.g. webservice interfaces) this can be done without much effort.

3.5 Prevention of Ad Fraud

Ad Fraud is the deliberate usage of an internet advertising system with the aim to impair one or more advertisers of that platform. The most prominent types of ad fraud stem from the payment scheme used: when

applying the "cost per mille" (CPM) schema an ad is shown on a website (e.g. as a "banner") and the advertiser has to pay a certain amount of money for each time the ad is shown. In the case of the "cost per click" (CPC) scheme the advertiser has only to pay when the user clicks on an ad, see Google (2005) for example. To perform ad fraud a malicious businessman might generate page impression or clicks on adverts of a competitor. Using special software these faked requests can be generated without much effort. For CPC usually a certain budget for one day is defined, so an ad fraud attack can burn up the budget of an advertiser and thus make his ads to disappear (Bomhardt, 2005). There are estimates that 10 percent or even more for particular businesses of clicks are fraudulent (Olsen, 2004) and taking into account that over 40 % of the 9.6 billion US $ of internet advertising revenues stem from performance based advertising like CPC (IAB, 2005) the effect of ad fraud isn't negligible.

In the short term the operator of the advertising platform is not impaired by ad fraud because contacts which stem from fraudulent actions also generated revenue. But in the medium or longer term the advertisers will find out that spending money on that advertising platform doesn't cause the intended effect (in most sales promotion) and thus ask the operator for a refund or stop advertising with that provider at all.

As discussed in chapter 2, mobile advertising can be seen as an advancement of internet advertising and may thus also be vulnerable to ad fraud, especially when the advertisers have to pay for each contact. Since this is the case for the MoMa-system, we also considered different measurements for the prevention of "mobile ad fraud". Ad fraud for MoMa would mean to sign up for many of the free user accounts and to create orders which fit an offer of the advertiser to be attacked. As we have to assume that the protocol used for the communication between MoMa-client and server is public, an attacker could develop a computer program to automate this.

Since ad fraud is mainly performed by programs or scripts which generate the fraudulent contacts an obvious approach for prevention would be the usage of so called CAPTCHAs. A CAPTCHA is an automated test to tell humans and computers apart by showing a little riddle which at the current state-of-the-art of artificial intelligence can only be solved by humans, e.g. an users sees a picture with distorted text and has to type in that text into an input field (von Ahn et al., 2004). Many providers of web-accounts use CAPTCHAs to prevent the automated registration of multiple account, but for advertising applications you cannot ask the consumer to solve a CAPTCHA before you grant him access to advertising information, especially not on a mobile device with its limited interface where the user should have to make as little as possible data entries.

To secure MoMa against ad fraud it shouldn't be possible to obtain an user-ID without being a human; it might even be necessary to perform some kind of age verification if there are advertising categories with content not suitable for minors. For each ID it is quiet simple to monitor if there is a conspicuous accumulation of orders which match a certain offer and to suspend the account with a given ID or restrict the number of orders for each category. This check has to be performed by the trustworthy party, because the matching-server only sees transaction pseudonyms and thus cannot determine if a set of orders originates from the same user.

4. SUMMARY

We discussed the features of mobile and wireless advertising as advancement of internet advertising. The introduced MoMa-system is based on the idea of permission marketing to comply with legal requirements and gain user acceptance. MoMa gives consideration to the special needs of a mobile application in several ways: because mobile terminals offer a limited user interface MoMa analyzes private and public context information and employs profiles to expect as little as possible entries from the user. Different notification channels are supported.

Since mobile terminals are very personal communication devices great importance was attached to technical measurements for guaranteeing data protection while also realizing a high degree of personalization. Spam is practical impossible as the advertisers don't have access to the end addresses; if an user doesn't want to receive advertising from MoMa he just has to suspend or delete active orders.

A special feature of the MoMa-context model is the discrimination of public and private context parameters. Unlike for public context parameters (e.g. weather) for the retrieval of private context parameters (e.g. current location of the user) access to the user's mobile terminal is required.

The members from industry of the MoMa-consortium plan to develop a mobile guide for soccer fans based on the results of the project.

ACKNOWLEDGMENTS

The project "MoMa — Mobile Marketing" and this work have been funded by the Federal Ministry of Economics and Labour, Germany (BMWA, contract no 01 MD 243); the responsibility for the content of this article lies solely with the authors. We would like to thank Jan Zeman and

Marti Bayo-Alemany for discussing technical details of the MoMa-system during the preparation of this paper.

References

Aalto, L., Göthlin, N., Korhonenm, J., and Ojala, T., 2004, A Bluetooth and WAP Push based location-aware mobile advertising system, in: *Proceedings of the 2nd international Conference on Mobile Systems, Applications and Services (MobiSys '04)*, Boston, USA, ACM Press.

Bauer, H., Lippert, I., Reichardt, T., and Neumann, M., 2005, *Effective Mobile Marketing*, Institut für marktorientierte Unternehmensführung, University of Mannheim, Germany.

Barkhuss, L., and Dey, A., 2003, Location-based service for mobile telephony: a study of users' privacy concerns, in: *Proceedings of the 9th IFIP International Conference on Human-Computer Interaction (Interact 2003)*, Zurich, Switzerland.

Barwise, P, and Strong, C, 2002, Permission-based mobile advertising, *Journal of interactive Marketing*, 16(1):14-24.

bmd wireless, and University of St. Gallen, 2005, First empirical global spam study indicates more than 80 percent of mobile phone users receive spam, http://www.mobilespam.org, visited at August 15, 2005.

Bomhardt, C., 2005, *Cheating on Online Ads*, Institute for Marketing at University of Karlsruhe (TH), Germany, Discussion Paper.

Chen, G., and Kotz, D., 2000, A Survey of Context-Aware Mobile Computing Research, Technical Report TR2000-381 of Dartmouth College, Hanover, NH, USA.

Genesereth, M., 1994, Software agents, *Communication of the ACM*, 38(7):48ff.

Godin, S., 1999, *Permission Marketing: Turning strangers into friends*, and friends into customers, Simon and Schuster, New York, 1999.

Google, 2005, Google AdWords, http://adwords.google.com, visited at August 12, 2005

Helm, S., 2000, Viral Marketing — Establishing Customer Relationships by "Word-of-mouse", *Electronic Markets*, 10(30):158-161.

Informa Telecoms & Media, 2005, *Global Mobile Forecast 5th Edition*, http://www.telecoms.com/gmforecasts, last visited at August 11, 2005.

IAB (Internet Advertising Bureau), 2005, *IAB Internet Advertising Revenue Report 2004*, http://www.iab.net/resources/adrevenue/pdf/IAB_PwC_2004full.pdf.

Kavassalis, P., Spyropoulou, S., Drossos, D., Mitrokostas, E., Gikas, G., and Hatzistamatiou, A., (2002) Mobile Permission Marketing: Framing the Market Inquiry, *International Journal of electronic Commerce*, 8(1):55-79.

Kölmel, B., and Alexakis, S., 2002, Location based Advertising, in: *Proceedings of the 1st International Conference on Mobile Business*, Athens, Greece.

Kotler, P., and Bliemel, F., 1992, *Marketing Management*, Poeschel, Stuttgart, Germany.

Krishnamurthy, S., 2001, A comprehensive Analysis of Permission Marketing, *Journal of computer mediated communication*, Volume 6(2).

MessageLabs, 2004, *Intelligence Annual Email Security report*, http://www.messagelabs.com

OECD, 2004, Organisation for Economic Co-operation and Development, *Background paper for the OECD Workshop on Spam.*

Olsen, S., 2004, *Google's fraud squad battles phantom clicks*, ZDNet Australia.

Mohr, R., Nösekabel, H., and Keber, T., 2003, V-Card: Sublimated Message and Lifestyle Services for the Mobile Mass-Market, in: *Proceedings of the 5th International Conference on Information and Web-Based Applications & Services*, Jakarta, Indonesia.

Netsize, 2005, *Netsize Guide 2005*, Paris, France; http://www.netsize.com

Pfitzmann, A., and Köhntopp, M., 2000, Anonymity, unobservability, and pseudonymity: A proposal for terminology, in: *Designing privacy enhancing technologies: International workshop on design issues in anonymity and unobservability*, Berkley, USA; Springer, Heidelberg, Germany.

Ratsimor, O., Finin, T., Joshi, A., and Yesha, Y., 2003, eNcentive: A Framework for Intelligent Marketing in Mobile Peer-To-Peer Environments, in: *Proceedings of the 5th international Conference on Electronic Commerce*, Pittsburgh, Pennsylvania, USA.

Schilit, B.N., Adams, N.I., and Want, R., 1994, Context-Aware computing applications, in: *Proceedings of the IEEE Workshop on mobile Computing Systems and Applications*, Santa Cruz, CA, USA, pages 85-90.

Schlickum, F., 2005, Erfahrungen in der Applikationsentwicklung mit J2ME, in: *Perspektiven des Mobile Business — Wissenschaft und Praxis im Dialog*, DUV, Wiesbaden, Germany.

Schwarz, T., 2001, Permission Marketing im Mobile Commerce, in: *Mobile Commerce — Grundlagen, Geschäftsmodelle, Erfolgsfaktoren*, Gabler, Wiesbaden, Germany.

Straub, T., and Heinemann, A., 2004, An Anonymous Bonus Point System for Mobile Commerce based on Word-Of-Mouth-Recommendation, in: *Proceedings of the 2004 ACM Symposium on Applied Computing*, Nicosia, Cyprus.

Toh, C.-K., 2002, *Ad hoc wireless networks: Protocols and Systems*, Prentice Hall, Upper Saddle River, NJ, USA.

Turowski, K., and Pousttchi, K., 2004, *Mobile Commerce*, Springer, Heidelberg, Germany

Ververidis, C., and Polyzos, G., 2002, Mobile Marketing using a location based Service, in: *Proceedings of the 1st International Conference on Mobile Business*, Athens, Greece.

Von Ahn, L., Blum, M., and Langford, J., 2004, Telling Humans and Computers apart automatically, *Communications of the ACM*, 47(2):57-60.

Wang, Z., 2003, *An agent based integrated service platform for wireless and mobile environments*, Shaker Verlag, Aachen, PhD-Thesis at University of Karlsruhe (TH), Germany.

Wohlfahrt, J., 2001, Wireless Advertising, in: *Mobile Commerce — Grundlagen, Geschäftsmodelle, Erfolgsfaktoren*, Gabler, Wiesbaden, Germany.

WURFL, 2005, Wireless Universal Resource File, http://wurfl.sourceforge.net, visited at August 23, 2005.

Yunos, H., Gao, J., and Shim, S., 2003, Wireless advertising's challenges and opportunities, *IEEE Computer*, 36(5):30-37.

Zeimpekis, V., Giaglis, G., and Lekakos, G., 2003, A Taxonomy of Indoor and Outdoor Positioning Techniques for Mobile Location Services. *SIGecom Exchanges*, 3(4):19-27, ACM Press, New York, NY, USA.

PRIVACY CHALLENGES FOR LOCATION AWARE TECHNOLOGIES

Carl Adams and Vasilios Katos
School of Computing, University of Portsmouth, PO1 3AE, UK

Abstract: Location aware capabilities can supply context and location sensitive information and support enabling users to be contactable and locatable within a wider mobile environment. These location awareness attributes can also be used to monitor user activities and movement through space and time. This paper explores location aware technologies and the resulting changing privacy and security landscapes for such mobile systems. The paper argues that the real challenge of meeting privacy obligations will be how to limit the joining-up or collaboration between the different monitoring technologies. However, this joining up capability is the very nature of information systems.

Keywords: Mobile Information Systems, Privacy, Security, Location Aware Technology

1. INTRODUCTION

A trend facing business and technology arenas is the move towards ubiquitous wireless technologies, which is having considerable strategic impact, fundamentally changing business models and processes (Barnes 2003,p2). Mobile information systems have location aware capabilities, providing context and location sensitive support. Some location awareness is also needed to provide access to the corporate infrastructure and systems, enabling users to be contactable and locatable within a wider environment. This paper explores location aware attributes and the implications of combining data from different sources. In addition the paper examines how the use and functionality of technology changes over time as users' needs change, using CCTV technologies as an example. Development of information systems with mobile capabilities has to recognise these location

aware attributes and the evolving nature of technology. There are clearly wider implications for business, and for developing mobile systems.

2. MONITORING FUNCTIONS OF TECHNOLOGY

Increasingly technology is used to monitor peoples' activity. The majority of the monitoring activities have been introduced for sound reasons such as ensuring the safety of pedestrians or protecting shoppers against fraud. However, monitoring activities raises privacy issues. Increasing monitoring activity has been the result of changes in business, technology and society. This has been most evident with CCTV. The UK is a 'good' example of this trend, as Privacy International (1995) identifies "[there has been an] extraordinary growth of the electronic visual surveillance industry ... In recent years, the use of Closed Circuit Television (CCTV) in the UK has grown to unprecedented levels". The number of CCTVs in the UK has grown dramatically since the mid 1990's, one estimate puts the number close to 4M (Lott 2005, p39). There are more CCTVs than the number of people able to sensibly monitor them. This clearly points towards automation and digitising of CCTV images. An example of wider use of such CCTV technologies is the introduction of congestion charging around London. This CCTV monitoring activity raises several privacy protection issues, as Spy.org (2005) identify: "We are not so much concerned with the actual congestion charge itself, but rather with the Privacy implications of the particular way in which this scheme has been implemented. ... The scheme relies on about 700 CCTV cameras covering around 203 entrances/exits to the 21 square kilometre central zone. Each lane of traffic either entering or leaving the boundary of the zone is covered by a [camera] linked to a Automatic Number Plate Recognition system".

The technology developed for congestion charging is now being used for other purposes. As The Register (2003) identifies, initially Ken Livingstone claimed that the system would only be used for congestion charging. However, only a short time after being introduced Ken indicated that the system can be used for other functions with cameras able to view drivers' faces, be controlled remotely and having variable angle and zoom facilities. In addition, the system would be used (when needed) to assist law enforcement activity.

Technologies evolve in how they are used: This concept is not new to most developers who have to contend with changing and evolving requirements. So a CCTV camera can be installed in a shopping space initially under the guise of addressing security issues, but evolves into a store management tool monitoring customer flows and purchase patterns.

initially under the guise of addressing security issues, but evolves into a store management tool monitoring customer flows and purchase patterns.

3. SECURITY AS THE MAIN DRIVER FOR MONITORING ACTIVITY

There are several drivers for increasing monitoring activity, not least because of an increased focus on security. Societal changes to be more security conscious has been recognised internationally in the OECD guidelines on promoting a global 'Culture of Security' (OECD 2003). The OECD's guidelines respond to the ever-changing nature of security and the wider environment, recognising the different roles and responsibilities for each 'participant' including Governments, businesses and society. The implications for owners and operators of information systems and networks are explicit in that they are expected to address the principles of risk assessment, and security design and implementation (ibid, p6). These changes in security focus, of course, have been influenced by events and changes in society, such as the 9/11 terrorist attacks in the US. In this global 'culture of security' Governments and corporations are expected to address security with appropriate policies and action, usually resulting in increasing collection and monitoring of information about employees and customers.

3.1 Data Protection and Privacy Issues

The increasing information collecting and storage capabilities of ICT have also resulted in increasing concerns over privacy issues. In the 1980's the OECD also developed a set of guidelines governing the protection of privacy (OECD 1980, p8). The guidelines identify a need for balance between privacy and economic needs to use, store and transmit such personal data to conduct transactions and support commerce. The guidelines identify basic principles for applying privacy protection laws, key of which is ensuring individuals have the opportunity of informed consent in the use of data about them. Governments and businesses have to achieve a somewhat difficult balance between meeting their security obligations on one-hand and privacy obligations on the other.

3.2 Implications of Mobile Location Aware Technologies

The potential for monitoring is likely to increase with the trends toward more mobile location aware technologies. Raina and Harsh (2002) describe

boundaries. This expansion of corporate space will also have a direct impact on privacy as information is transmitted over this wireless corporate space: "As a corporation's corporate space expands it may also overlap with other corporation's corporate space. The overlapping may result in and clashing of corporate spaces, or development of multi corporate wireless spaces where more than one business is operating and sharing mobile information" (Adams and Katos 2005). This expanded corporate or multi-corporate spaces may also encroach on individuals' personal space, or personal trust space (Adams et al 2003).

From a security perspective wireless technologies are inherently less secure opening the system up to attacks from outside the corporation's premises and providing an increase in potential security weak points (Panko 2004;Ciampa 2004, p22). Wireless security can be improved to reach (something equivalent to) the wired world but as Ciampa (2004, p23) identifies extra technologies and processes will be required. Increasingly wireless involves not just one mobile technology but an array of competing devices and infrastructure each with their own security (and privacy) issues.

From privacy perspective wireless technologies also raise more concerns as location information is embedded in the protocols or within system attributes. Many of the wireless devices are personal devices (e.g. PDAs, Bluetooth phones) attached to individuals and so likely contain personal information. They will also be locatable within an organisation space by association with their access points and interaction with other devices. It may even be possible to monitor movement of such devices, and their users, throughout the corporate space. The same monitoring capabilities are possible outside the corporate space. For instance, since 1996 wireless carriers in the US have been required to incorporate wireless phone locators in their networks for 'safety' reasons (WLIA 2001b) under the E911 (Enhance 9.1.1. the emergency services telephone number in the US) mandate as part of the Telecommunications Act of 1996. Similar early location based developments took place in Japan and Europe (ibid). Most wireless telecommunications infrastructures incorporate location information about users, embedded in operating protocols. With 3G technologies and always-on capabilities, such location information will become more accurate, pervasive and intrusive.

The importance of privacy and anonymity with location aware technologies has been recognised by the Wireless Location Industry Association (WLIA): "Without question, the single most important issue confronting the new industry of wireless signal location technology and applications is the issue of personal privacy. As with any other issue, there are two sides to this issue. This does not mean that there are two sides as in "for" or "against" privacy, nor two sides as in "for" or "against" wireless

Association (WLIA): "Without question, the single most important issue confronting the new industry of wireless signal location technology and applications is the issue of personal privacy. As with any other issue, there are two sides to this issue. This does not mean that there are two sides as in "for" or "against" privacy, nor two sides as in "for" or "against" wireless signal location. Reasonable people will look for solutions that will maintain both the value of protecting individual privacy and the value of achieving the benefits of new wireless location technologies. The issue is how to strike the right balance between these values." (WLIA 2001a)

The WLIA have been involved in developing a set of self regulation policies for operators involved in location monitoring activity, these include the Fair Location Information Practice (FLIP) (Barnes 2003, p144; WLIA 2001a), key of which is also ensuring individuals have the opportunity of informed consent in the use of location data about themselves. However, there are practical challenges in applying the FLIP guidelines. For instance, most of the data collection activity is likely to be automatic. Even if some informed consent activity is possible, it is unlikely to be fully exercised since much of the data will to be collected and stored under security and monitoring obligations: Telecommunication operators and providers have to maintain logs of customer call activity, including location information (Adams and Katos 2005).

When more than one technology interoperates, the privacy space may shrink substantially. For example, a CRM related technology employed by a supermarket might use customer loyalty cards. The purpose of these cards is to collect valuable marketing information related to buying patterns; the customer willingly sacrifices a 'small amount' of privacy for a minuscule financial gain. However, even with customers without loyalty cards, hypothetically at least, companies are still able to populate their CRM database by combining transaction information with CCTV or with information broadcasted by any wireless personal devices a customer may have (bluetooth, RFIDS etc). This raises a more fundamental privacy challenge. The privacy landscapes are likely to change considerably once there is 'joining up' between monitoring technologies. Combining data from more than one source enables a richer set of data to be collected about individuals. All the monitored data can be stored electronically, possibly on the same computer systems. All the technology seems to be in place to bring together the different monitoring activity: Not only does a corporation have the capability to track what you buy but also how you entered the purchasing space and all the different activity in moving through that space right though to the purchasing. A psychologist analysing this information may be able to gleam considerably more information about individuals than the individuals may be aware of themselves!

4. EVOLVING USE OF LOCATION AWARE TECHNOLOGIES

As seen with the CCTV examples above, many of the CCTV monitoring technologies seem to fall outside the realms of data protection and privacy rules and controls. Equally there seems little to limit expanding the function and scope of using such technologies as well as collating and collaborating monitoring information from different sources.

The de facto approach in collection of location information seems to automatically imply consent to collecting and using information in what ever fashion the collectors' wish. The possibility of collaboration between location aware technologies is, at least in part, also likely to be automatic and location based services will require some joining up of information.

Take the example of the possible introduction of ID cards in the UK. ID cards are used in a variety of other countries, however, the proposed application in the UK will result in generating a vast repository of personal data on individuals in the UK, including biometrics (LSE 2005). The ID cards are likely to - at least in part - be in the form of contactless cards. During 2004 the proposed national ID card Bill had a rough ride through parliament and was not able to make it through the House of Lords before the call for new elections in the UK in 2005 (Privacy International 2005). However, very soon after the UK election in early May 2005, the continuing Labour Government reintroduced The Identity Cards Bill on the 25 May 2005 (UK Gov 2005). The UK Government seems serious about 'pushing through' a national biometric based ID card system. The same is true in the US where technological developments towards wireless ID cards is more advanced, along with the 'pushing through' of legislation. The US ID card bill in the US seems to have been passed through the senate on the back of an $83Bn funding bill for troops in Iraq and general homeland security (Privacy International 2005). In both the UK and US a rich set of personal and biometric information will be communicated over a wireless medium. Presumably there will be some demand to people to carry their electronic ID cards. A uniquely identifiable ID card with electronic communicating capabilities opens up a range of possible uses and more particularly for implementing authentication services. It could also be used for a range of other functions such as a driving licence, access to a toll both or corporate space or as the basis for facilitating electronic interactions such as e-voting. It is likely that many types of organisations will be interested in using such 'useful' devices and it would seem a natural progression to allow this. Commercial avenues may be an attractive proposition for governments that have spent the estimated tens of billion £'s taxpayers' money to develop the infrastructure (LSE 2005). Mobile information systems for such

organisations are likely to have access to very personal information, including the location and context relative information from any interaction with individuals.

With such an array of different wireless technologies ubiquitously available, including RFIDs, Bluetooth, WiFi, GPS and mobile phone technologies, then it seems a natural progression that corporations will be developing data contact and collection points with these. The use for these contact and collection points is also likely to evolve. The range and volume of personal data to be collected is growing and corporations' mobile information systems will be playing a part in managing this data and making sense of it. Making sense of vast amounts of information is what information systems do, and do remarkable well. Mobile information systems will also play a part in the natural evolution of the use of mobile technologies and seems a natural outcome of the 'making sense' process. The scenario described earlier of corporations being able to 'gleam considerably more information about individuals, more than the individuals may be aware of themselves' seems a natural outcome of the progression towards ubiquitous wireless systems.

5. CONCLUSION

This paper has contented that there has been an increase in the use of technology to monitor peoples' activity, and that there is increased potential for monitoring with the trend towards more mobile location aware technologies. Location aware capabilities offer potential benefits to customer and citizens, but equally they raise privacy issues. Individually the privacy issues are not insurmountable. However, as the use of such technologies unfold and develop, coupled with the capabilities of combining location information from different channels there are considerable impacts on privacy and anonymity. The trend towards location aware technologies seems inevitable. Also inevitable is an increase in the amount of location information collected, given that it is most automatic and part of the technology operation. The operation of the wireless technologies also makes joining up of monitoring activity more likely; indeed it will be a prerequisite for providing many location aware services.

This paper discussed the developing privacy landscapes, particularly with reference to mobile location aware technologies. The real challenge of protecting citizens' privacy will be how to limit the joining-up or collaboration between the different monitoring technologies, and this joining up capability and making sense of such joined up data is the very nature of information system. A real challenge for future mobile information systems

is how to protect privacy while managing a vast and growing set of location aware and, increasingly personal data.

References

Adams C., Avison D.E. and Millard P. (2003) Personal trust space in mobile commerce. Proceedings of ICECR-6, Dallas, Texas, Oct 23-26, 2003, p396–403.

Adams C. and Katos V. (2005) The ubiquitous mobile and location aware technologies time bomb. Cutter IT Journal, June 2005.

Barnes S. (2003) Mbusiness: The strategic implications of wireless communications. Elsevier Butterworth-Heinmann, London.

Brunk, B. (2003). A Framework for Understanding the Privacy Space. PhD Thesis, University of North Carolina.

Ciampa M. (2002) Guide to Wireless Communications. Thomson, Massachusetts.

Crimereduction.gov (2005) CCTV initiatives home page. A UK government, pro-CCTV web site. http://www.crimereduction.gov.uk/cctvminisite4.htm

HEW (1973). Secretary's Advisory Committee on Automated Personal Data Systems, Records, Computers, and the Rights of Citizens (HEW Report). U.S. Department of Health, Education and Welfare.

Lott T. (2005) Every move you make. Times newspaper, Magazine, May 14th, p37-41.

LSE (2005) The Identity Projects: An assessment of the UK Identity Cards Bill and its implications. Version 1.09, June 27, 2005.London School of Economics.

Martin L. (2005) This chip makes sure you always buy you round. The Observer, 16th January 2005, p14.

OECD (1980) OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data. OECD Publications, Paris. Also available from the web at www.oecd.org

OECD (2003) Implementation plan for the OECD guidelines for the security of information systems and networks: Towards a culture of security. Document DSTI/ICCP/REG(2003)/5/REV1, 02-July-2003, Organisation for Economic Co-operation and Development.

Olsen, S. (2000). Web browser offers incognito surfing. CNET News.com, http://news.com.com/ Web+browser+offers+incognito+surfing/2100-1017_3-247263.html

Panko R.R. (2004) Corporate Computer Network Security. Prentice-Hall, NJ.

Privacy International (2005) "UK ID Card Bill Dies" Available from http://www.privacyinternational.org/article.shtml?cmd[347]=x-347-178984

Raina K. and Harsh A. (2002) mCommerce Security: A beginner's guide. McGraw-Hill/Osborne, California.

Schneier, B. (2000). Secrets & Lies. New York: John Wiley & Sons.

Spy.org (2005) London Congestion Charge CCTV privacy concerns. From http://www.spy.org.uk/cgi-bin/cclondon.pl

The Register (2003) "London charge zone is security cordon too, says mayor" from http://www.theregister.co.uk/2003/02/17/london_charge_zone_is_security/

UK Gov (2005) Identity Cards Bill. http://www.publications.parliament.uk/pa/cm200506/cmbills/009/2006009.htm

Warren S. and Brandies, L. (1890). The Right to Privacy. Harvard Law Review 4(5)

WLIA (2001a) DRAFT WLIA PRIVACY POLICY STANDARDS. Available from the web at http://www.wliaonline.org/publications/.

WLIA (2001b) A DIALOGUE ON PRIVACY ISSUES AND WIRELESS LOCATION SERVICES. Available from http://www.wliaonline.org/publications/argue.html.

SEEKING ANSWERS TO THE ADVANCED MOBILE SERVICES PARADOX

Minimal Acceptance and Use Despite Accessibility

Jennifer Blechar[1], Ioanna Constantiou[2], Jan Damsgaard[2]
[1]*University of Oslo, Norway; Copenhagen Business School, Denmark*

Abstract: At a time when mobile service adoption rates remain much lower than anticipated, better understanding of users' behavior and attitude towards new technology becomes critical. This paper brings insights from behavioral economics to the technology acceptance research domain to investigate adoption of advanced mobile services. Through a field study of mobile service acceptance and use, we explore user behavior during the initial stages of technology adoption. We argue that in order to better understand the cognitive processes underlying technology acceptance or rejection decisions for advanced mobile services, the economic aspects of those processes must be considered. This is especially the case given the similarity of new mobile services to those currently available to consumers through other means and the inherent elements of substitution which exist therein.

Keywords: Technology Acceptance, Mobile Services, Mental Accounting

1. INTRODUCTION

As the prices on voice-based telephony steadily decrease, the mobile telecommunications industry is looking for new sources of revenue. Several of the key actors in the field have pinned their hopes on advanced mobile services such as news, weather forecasts, chat and map directories offered via mobile operators' proprietary platforms. However, in many western European countries mobile users have been reluctant to embrace these services let alone pay for them, while at the same time, sales of new mobile devices are increasing at an unprecedented pace. There seems to be a paradox; the new mobile devices can support various advanced services but

users do not take the step of initiating and adopting them. In this paper we seek to better understand why users do not adopt new mobile services even when they are potentially readily available in the palm of their hand.

As many technologies mature they become domesticated and find their place in our everyday lives. For example the radio found its place, consumed a certain part of our time and became an important ingredient in our lives. When the TV came around in 1950's it challenged the position of the radio and part of the time dedicated to radio was substituted by TV. In the recent decade we have seen time spent on Internet related activities increasing at the expenses of the TV. Thus, with the adoption of a technology, use substitutes time spent on other activities. In our context, new advanced mobile services must challenge, compete and replace traditional mechanisms of getting access to news, weather forecasts, etc. It is this process we have not yet seen progress for in the mobile services arena.

In this paper we introduce the concept of mental accounting from the behavioral economics domain to technology acceptance in the mobile industry as an explanatory vehicle to investigate this lack of adoption of advanced mobile services. Mental accounting addresses the cognitive processes underlying users purchase and use decisions and offers the opportunity for a systematic investigation of choice processes. The next section briefly reviews existing related literature in the mobile services arena and introduces the concept of mental accounting. This is followed by an introduction to the conducted field study of new mobile service use among 36 mobile service users in Denmark. The discussion exploring the preliminary findings through the lens of mental accounting is then presented. Finally, we point to main insights generated and future research directions.

2. LITERATURE INSIGHTS

2.1 Technology Acceptance in the Mobile Arena

The study of technology adoption continues to be a popular topic of concern within the IS discipline. Traditional research in this domain is rooted in models and theories of technology acceptance such as the Technology Acceptance Model (Davis, 1989), The Theory of Planned Behavior (Ajzen, 1985, 1991), the Theory of Reasoned Action (Ajzen & Fishbein, 1980), etc. As research within the mobile industry progresses, many of the models above have also been adopted and used to investigate the acceptance of new mobile services and other mobile applications (e.g. Hung et al., 2003; Khalifa & Cheng, 2002).

However, many of the services offered through mobile devices are available through existing technology, such as the Internet, thus evoking an element of substitution. This is an aspect not necessarily present in existing technology adoption studies where the above models have typically been applied. Many of the mobile services being offered today are free Internet services that have simply been transferred to the mobile device resulting in poorer quality with an expectation that the consumer will now be willing to pay for them. Finally many services only have value when a majority has adopted them due to network externalities (Katz & Shapiro, 1992).

In order to more thoroughly explore technology acceptance of new mobile services, we argue, existing research must consider the economic aspect of technology adoption within the wider context of use. Choosing to pay for a novel mobile service also offered through the Internet, for example, can contain different motivational factors than making use of a required organizational program. In doing this, focus must be placed on the underlying cognitive elements fueling users choice and decision processes as well as the ongoing reflexivity present during the technology acceptance process (Blechar et al., 2005). As such, insights from the economics discipline seem to offer promise for the further exploration of potentially influential elements related to user choice and decision processes.

2.2 Mental Accounting

The concept of *mental accounting* comes from the behavioral economics domain and deals with the cognitive processes underlying consumer choices and decision behaviors (Thaler, 1980). It embarks from the neo-classical economic models by combining insights from psychology to the analysis of consumer's choices, thereby offering complimentary insights to existing technology acceptance research. It includes four main components that explain; how users perceive and experience the outcome of choice processes, the underlying cognitive processes influencing purchase decisions, the manner in which income is allocated to the purchase of various goods and the frequency with which consumers evaluate their income versus purchases (p. 183, Thaler, 1999). This is particularly useful in attempting to explain consumers' motives and decisions related to adoption of different products or services.

Thaler (1985) proposed a set of parameters for mental accounting. We focus on the two parameters related to the mobile services context that offer insight on mobile users' behavior, namely, *transaction utility* and *payment decoupling*. "Transaction utility measures the perceived value of a deal. It is defined as the difference between the amount paid and the 'reference price' for the product/service" (p. 188-189, Thaler, 1999). The latter is the

"regular" price that the consumer expects to pay for this product/service according to the reference he makes in his mind to a similar situation or experience of purchase. Payment decoupling separates the purchase of a product from its consumption in the mind of the consumer (p. 192, Thaler, 1999). This separation seems to reduce the perceived cost of the purchase and can be achieved through pre- (e.g. prepaid cards) or post- (e.g. credit cards) payment.

3. THE FIELD STUDY

This paper presents preliminary results of a field study of new mobile service use and adoption in Denmark. 36 mobile phone users from three categories: students at an affiliate university, employees at a public agency and employees within the Department of Informatics within our University, were included in the study. The average age of participants was 31 and 40% were females. Users were given new mobile phones and pre-paid SIM cards providing access to a wide variety of GPRS/WAP enabled content services. These services, available through the participating mobile operators' portal, provided access to information such as news, weather, local information and chatting functions. The purpose of the study was to observe the behavior of new mobile service users and the manner in which they accepted or rejected the use of the mobile services included in the study.

The project followed a mixed methods approach (Creswell, 2003; Hammersley, 1996) and thus several qualitative and quantitative data was collected throughout the project period, which lasted from November 2004 to March 2005. The most prevalent form of data collection was through surveys distributed to participants at various points throughout the project. Surveys included open ended and fixed response questions, based primarily on criteria from the combined UTAUT model of technology acceptance (Venkatesh et al., 2003). Data collection also included short interviews conducted with 17 participants at the close of the project. These followed an interview guide and were recorded and semi transcribed.

Once collected, data was analyzed through various means. From the interviews, topics which were frequently raised by participants were identified and extracted based on the queries to participants. These were compared to the results from the open ended responses included on the surveys to distill overlying themes across the project. The final set of topics raised were explored and analyzed according to the concepts of utility and pricing. In addition, participants' selections to the fixed response questions on the surveys were gathered and analyzed through a commercially available statistical survey analysis tool.

3.1 Results and Discussion

We introduce the concept of mental accounting to elaborate on participants behaviors and attitudes towards mobile services based on the data collected from our field study. A key component of mental accounting is the consumers' reference price for the service under investigation. We observe from both interviews and survey data that participants related mobile services offered through the portal to the Internet services already available to them, as the following comment illustrates:

> "Because I have access to a computer most of the times and those services are just as easy even easier to use on a regular computer..." (Male, 25, Interview Response)

This is further exemplified by participants' response in the surveys where 71% of the participants agreed or strongly agreed that the new mobile services did not offer anything new to what was already available to them through the Internet. The corresponding Internet services (i.e. news, weather forecast, map directories, chatting etc.) were typically perceived by participants as being free of charge. This could be explained by the fact that participants either had access to these services from work or school where their employer was paying for the Internet access or from home where the payment of the services is decoupled from the fixed monthly charges they pay for their Internet access.

This had direct implications on participants' reference price for many of the mobile services offered through the mobile portal. In particular participants seemed to have a reference price that was minimal, close to zero. As one participant noted:

> "... The problem with the mobile services is, that I have to pay for information, that has become commodities. Why pay for news, which is free on the internet? Why pay 2kr for a joke, when I can get thousands for free on the internet? I need something original and unique for a mobile service." (Male, 27, Survey Response)

Further, when participants were queried regarding the value of the new mobile services, 57% indicated that the mobile services available did not add value to their everyday life.

Consequently in terms of transactional utility, it seems that there was a negative effect in the use of mobile services because of their perceived high price. In particular, participants' perceived mobile services as expensive as compared to the perceived free of charge Internet services. Besides in terms of acquisitional utility, the mobile services were viewed as providing limited added value, generating low satisfaction because of poorer quality and speed

as compared to those services already available to participants through other means:

> "What the problem is that its like they have tried to take something from the Internet and put it down on a small screen instead of actually thinking, OK, what can we use this small screen for that is different than what is already on the Internet because when you can get something that you can get so much better while sitting with your laptop or something then there is no need sitting there with your phone, and waiting for it to come because its also very slow and the quality is not the same on that small screen" (Female, 26, Interview Response)

In turn these negative effects from both transactional and acquisitional utilities affected the participants' willingness to pay for the mobile services. Indeed, 62% of the participants indicated that the prices would be a barrier to the adoption of the mobile service offered.

Interestingly, the context of use of the mobile services was pointed to as an element impacting participants' willingness to adopt and use those services. Participants indicated that this context or physical location was one barrier to their mobile service use:

> "what I like to use my phone for is when I am not close to the computer or the internet..." (Female, 33, Interview Response)

However, Participants seemed willing to pay and make use of the mobile services in certain circumstances (such as when there was a power outage and they were unable to access information through existing means). As some participants noted:

> "I also tried to buy news and that was OK. But it was also a limited amount of news you got though and it was something I could see on my TV as well if I had text TV...one day I lost electricity in my house so I went on my phone instead and that kind of nice to have that opportunity" (Female, 26, Interview Response)

In these situations, participants did not have the ability to compare mobile services to the payment decoupled Internet services they typically had access to from work or their home environment. Therefore the mobile services use was perceived as value adding. It appears that the services offered through the mobile portal were only a 'back-up' when their typical means of access failed, or 'for fun' when they were traveling, etc. This indicates that the new mobile services need to be distinguished in some way, or be made to 'fit' users existing cognitive frame of how the services should function, what information they should provide and at what cost.

4. CONCLUSION

While many consumers are purchasing new mobile phones with capabilities to access and use advanced mobile services, adoption of those services is still dwindling. Through the concept of mental accounting from the behavioral economics research domain, this paper has presented preliminary evidence towards the exploration of why consumers have not adopted advanced mobile services despite their availability. This paper suggests that existing technology acceptance research in the mobile industry should be extended to consider the mental accounting present in users actions related to the utility, pricing and value of new mobile services.

While existing technology adoption research has to some extent investigated the cognitive processes of consumers in the ongoing process of technology consumption, advanced mobile services offer a context of use which differs from many of the previous studies. Namely, mobile services in their current form are essentially substitutes for services already available through other means, such as the Internet. Thus while mobile services do provide information which many users desire and are accustomed to access on a daily and regular bases (such as news or weather), the use of the mobile as a device and service mechanism has not yet been defined in the cognitive processes of the users. This indicates that mobile operators must first define this context if advanced mobile services are going to be successfully adopted and diffused in the market.

Future research in this domain should expand on the results presented in this paper to further explore the behavioral and cognitive elements of users actions motivated through economic elements of choice. As of yet, limited work has been done to join the economic discipline and IS adoption studies in the mobile arena. We believe that bringing insights from both these strands of research is critical for a fuller understanding of the process in which users make decisions related to the adoption of new technologies to their everyday lives. There are time limits in everyone's day and while the television was successful in converting radio listeners' time into television viewing time, the same phenomenon has not yet occurred for advanced mobile services. Better understanding the economic motivations behind this non-adoption is critical if mobile operators are to reap the benefits they expect to gain from full scale deployment of those services.

ACKNOWLEDGEMENT

This research was carried out in the realm of the Mobiconomy project at Copenhagen Business School. The research was in part supported by the Danish Research Agency, grant number #2054-03-0004.

References

Ajzen, I. 1985. From Intentions to Actions: A Theory of Planned Behavior. In J. Kuhl, & J. Beckman (Eds.), Action Control: From Cognition to Behavior. New York: Springer.

Ajzen, I. 1991. The Theory of Planned Behavior. Organizational Behavior and Human Decision Processes, 50: 179-211.

Ajzen, I., & Fishbein, M. 1980. Understanding Attitudes and Predicting Social Behavior. New Jersey: Prentice Hall.

Blechar, J., Knutsen, L., & Damsgaard, J. 2005. Reflexivity, the Social Actor and M-Service Domestication: Linking the Human, Technological, and Contextual. Paper presented at the IFIP TC8 WG 8.2 Working Conference on Designing Ubiquitous Information Environments: Socio-Technical Issues and Challenges, Cleveland, Ohio, USA.

Creswell, J. W. 2003. Research Design: Qualitative, Quantitative and Mixed Methods Approaches (second ed.). Thousand Oaks: Sage Publications.

Davis, F. D. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly, 13: 319-339.

Hammersley, M. 1996. The Relationship Between Qualitative and Quantitative Research: Paradigm Loyalty Versus Methodological Eclecticism. In J. T. E. Richardson (Ed.), Handbook of Qualitative Research Methods for Psychology and the Social Sciences: 159-174. London: The British Psychological Society.

Hung, S.-Y., ku, C.-Y., & Chang, C.-M. 2003. Critical Factors of WAP Service Adoption: An Emperical Study. Electronic Commerce Research and Applications, 2: 42-60.

Katz, M. L., & Shapiro, C. 1992. Product introduction with network externalities. Journal of Industrial Economics, 40(1): 55-83.

Khalifa, M., & Cheng, S. K. N. 2002. Adoption of Mobile Commerce: Role of Exposure. Paper presented at the 35th Hawaii International Conference on System Sciences (HICSS), Big Island, Hawaii.

Thaler, R. H. 1980. Towards a positive theory of consumer choice. Journal of Economic Behavior and Organization, 1: 39-60.

Thaler, R. H. 1985. Mental Accounting and Cosumer Choice. Marketing Science, 4: 199-214.

Thaler, R. H. 1999. Mental Accounting Matters. Journal of Behavioural Decision Making, 12: 183-206.

Venkatesh, V., Morris, M., Davis, G. B., & Davis, F. D. 2003. User Acceptance of Information Technology: Towards a Unified View. MIS Quarterly, 27(3): 425-478.