HAL
open science

Evaluation des systèmes mobiles et ubiquitaires: proposition de méthodologie et retours d'expérience

Francis Jambon, Nadine Mandran, Brigitte Meillon, Christian Perrot

To cite this version:

Francis Jambon, Nadine Mandran, Brigitte Meillon, Christian Perrot. Evaluation des systèmes mobiles et ubiquitaires: proposition de méthodologie et retours d'expérience. Journal d'Interaction Personne-Système, 2014, Volume 1 (1), pp.1-34. 10.46298/jips.61 . hal-01058933

HAL Id: hal-01058933

https://inria.hal.science/hal-01058933v1

Submitted on 28 Aug 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Évaluation des systèmes mobiles et ubiquitaires : proposition de méthodologie et retours d'expérience

FRANCIS JAMBON, NADINE MANDRAN,
BRIGITTE MEILLON ET CHRISTIAN PERROT
Laboratoire d'Informatique de Grenoble / équipe MultiCom
Université de Grenoble et UMR CNRS 5217

Résumé : L'évaluation des systèmes interactifs mobiles et/ou ubiquitaires par l'intermédiaire des tests utilisateurs semble a priori plus pertinente sur le terrain qu'en laboratoire d'utilisabilité. Pourtant, les résultats de la littérature apparaissent comme contradictoires. Notre objectif dans cet article est d'en expliciter les raisons et de proposer une méthodologie minimisant les biais. Les expérimentations décrites dans la littérature et nos propres travaux nous ont amenés à définir le concept d'environnement interactif et trois approches expérimentales possibles : laboratoire, terrain et situation réelle. Nous proposons ensuite une méthodologie et une technique (le cheval de Troie) adaptées à l'évaluation en situation réelle. Enfin, nous illustrons notre approche théorique par trois expérimentations et en donnons des retours d'expérience. Nous concluons ensuite sur les limites de notre approche.

Mots clés : Systèmes mobiles, informatique ubiquitaire, évaluation en situation réelle, technique du cheval de Troie.

Abstract: The evaluation of mobile and/or ubiquitous interactive systems via user testing seems a priori more relevant in the field than in a usability laboratory. However, the results of the literature are contradictory. In this article, we aim at explaining the reasons why, and we propose a methodology that could minimize biases. The experiments described in the literature and our own experiments lead us to define the interactive environment concept and three possible experimental approaches: laboratory, field and reality testing. Then, we propose a methodology and a technique –the Trojan horse– adapted to the evaluation in reality testing. At last, we illustrate the theoretical approach by three experiments and give experience feedbacks on them. We conclude on the limits of our approach.

Key words: Mobile devices, ubiquitous computing, reality testing, Trojan horse technique.

Les articles de JIPS sont publiés sous licence Creative Commons Paternité 2.0 Générique.

1. INTRODUCTION

Les systèmes mobiles ont franchi, il y a plusieurs années déjà, le seuil de la maturité technique et ont trouvé un réel usage auprès de nombreux utilisateurs. Parmi les dispositifs largement diffusés, citons par exemple le système TomTom® de géolocalisation et de routage pour véhicule ou les téléphones mobiles BlackBerry® permettant de lire le courrier électronique. Évolution de ces systèmes mobiles, les systèmes dits ubiquitaires, que l'on regroupe souvent sous le terme plus général d'intelligence ambiante, sortent progressivement de l'état de maquette de laboratoire pour être mis en œuvre dans le monde réel, par exemple dans les musées [Sermet & Millet, 2007].

Ces deux catégories de systèmes permettent la mobilité de l'utilisateur, que ce soit **avec** le système (cas des systèmes mobiles) ou **dans** le système (cas des systèmes ubiquitaires). En outre, le principal intérêt de ces systèmes est d'être sensibles à leur environnement (géolocalisation, disponibilité des réseaux de communication, identification de l'utilisateur, etc.). Ces deux caractéristiques, qui font l'essence même de ces systèmes, ouvrent de nouvelles perspectives concernant les méthodologies adaptées à leur évaluation [Kjeldskov & Graham, 2003], [Hagen, Robertson, Kan, & Sadler, 2005]. En effet, l'évaluation par l'intermédiaire des tests utilisateur de tels systèmes impose de se placer dans un ensemble de situations qui donne tout le sens à leur usage. Ce n'est donc pas seulement un « système interactif » que l'on doit évaluer, mais plus généralement un « environnement interactif ». Toute la difficulté est de définir puis créer cet environnement.

Dans cet article, nous explicitons la problématique et les principaux travaux du domaine qui nous ont amenés à l'évaluation des systèmes interactifs en situation réelle. Nous détaillons ensuite la méthodologie que nous proposons pour leur évaluation, ainsi que ses limites. Enfin, à la lumière des expérimentations que nous avons menées, nous décrivons notre retour d'expérience à la fois méthodologique et technique, puis concluons sur la faisabilité de telles expérimentations. Cet article constitue une synthèse revue et corrigée des précédentes publications de notre équipe sur ce thème : [Jambon, 2006, 2009; Jambon, Golanski, & Pommier, 2006, 2007; Jambon, Mandran, Meillon, & Perrot, 2008; Jambon, Mandran, & Perrot, 2007; Jambon & Meillon, 2009].

2. LIMITES DES APPROCHES ACTUELLES

Afin d'évaluer un système mobile ou un ensemble de systèmes participant à de l'intelligence ambiante en présence de l'utilisateur, il est nécessaire de placer cet utilisateur dans l'environnement interactif tout en gardant assez de contrôle et de capacité d'observation de cet environnement pour pouvoir analyser l'interaction entre l'utilisateur et le système ou l'ensemble de systèmes.

2.1. Environnement interactif

En premier lieu, il est nécessaire de déterminer quels sont les éléments de cet environnement qui seront pertinents vis-à-vis des objectifs de l'évaluation. Puis, en second lieu, il convient de définir comment l'utilisateur peut être placé en interrelation avec ces éléments, soit en les simulant, soit en faisant appel à des éléments issus du monde réel. Du point de vue de l'évaluation, nous proposons de regrouper les éléments de l'environnement interactif selon quatre catégories :

- **l'utilisateur** ou les utilisateurs dans le cas d'un environnement collaboratif ;

- le ou **les dispositifs** matériels et logiciels qui sont l'objet direct de l'évaluation ;
- **les tâches** liées aux fonctionnalités du ou des dispositifs que l'on souhaite évaluer ;
- **le contexte**, que nous définissons comme le complémentaire des trois éléments précédents : les autres utilisateurs, les autres dispositifs au sens large (qu'ils soient informatisés ou non) et l'activité de l'utilisateur sans rapport direct avec le dispositif évalué.

L'une des caractéristiques saillantes du contexte est qu'il peut varier en fonction du déplacement de l'utilisateur. De plus, faire la distinction entre ce qui est contextuel et ce qui ne l'est pas est parfois difficile, car par définition ces systèmes s'intègrent dans leur environnement. En effet, certains éléments du contexte participent directement à l'interaction, alors que d'autres font partie de « l'ambiance générale » et ont un rôle bien moindre, sans toutefois pouvoir être considérés comme négligeables. C'est par exemple le cas des conditions météorologiques dans une station de ski, de l'affluence dans un musée.

2.2. Laboratoire ou terrain ?

La question que l'on doit se poser est : quel degré de réalisme doit-on apporter au contexte de l'environnement interactif pour assurer la validité des analyses ? Faut-il privilégier le laboratoire ou le terrain ? Cette question n'est pas nouvelle en ergonomie. En effet, elle a été abordée dès les années soixante, notamment par Chapanis [Chapanis, 1967]. Dans le cadre des systèmes mobiles et ubiquitaires, l'importance de la mobilité, laquelle influençant la notion de contexte et sa variabilité, ainsi que les progrès techniques accomplis concernant les systèmes d'acquisition (permettant ainsi leur usage dans de nombreux contextes), nous ont incités à revisiter la question.

Dans la littérature, nous distinguons principalement deux approches. Une première approche consiste à simuler l'ensemble des éléments du contexte avec plus ou moins de réalisme. Il s'agit là d'expérimentations en laboratoire d'utilisabilité, dont la méthodologie est bien connue. Pour cette approche, l'un des aspects importants du protocole expérimental est le moyen utilisé pour simuler la mobilité de l'utilisateur. Une seconde approche consiste à placer l'utilisateur dans le monde réel, où les éléments du contexte sont déjà « naturellement » présents. Ce sont les évaluations dites « de terrain ». Pour ce second type d'approche, un élément important est le relatif éloignement de l'utilisateur par rapport aux observateurs, qui rend plus délicate l'acquisition des données expérimentales. Sur ce point, cette approche est similaire aux méthodes d'évaluation à distance.

Globalement, il pourrait sembler « évident » que les évaluations sur le terrain aient une plus grande validité expérimentale que celles effectuées en laboratoire, car nous pouvons être certains que tous les éléments attendus du contexte sont présents, du fait même que l'expérimentation se situe dans la réalité. Nous allons voir que la littérature n'est pas unanime à ce sujet.

Simulation de la mobilité

La mobilité de l'utilisateur peut être simulée de différentes manières. La simulation la plus simple consiste à faire se déplacer l'utilisateur selon un parcours prédéfini, souvent circulaire, autour d'obstacles, et dans un espace restreint : autour d'un bâtiment [Brewster & Walker, 2000], dans un couloir [Pirhonen, Brewster, & Holguin, 2002], ou

dans une salle d'expérimentation [Kjeldskov & Stage, 2004]. Une autre technique consiste à détourner de leur usage prévu (l'entrainement des sportifs) certains systèmes simulant la marche comme les mini-steppers [Pirhonen et al., 2002] ou les tapis motorisés [Kjeldskov & Stage, 2004]. Ces techniques permettent de recréer les mouvements et vibrations que peut ressentir un utilisateur en marchant, et ainsi tester notamment la lisibilité et la facilité avec laquelle il utilise les dispositifs d'interaction dans ces conditions.

Remarquons que la seconde technique est la moins contraignante pour l'instrumentation car elle n'oblige pas l'utilisateur à se déplacer physiquement. Cela permet d'utiliser des caméras fixes ainsi que divers câbles comme l'alimentation électrique du prototype ou le réseau. La première n'a pas ces facilités, mais requiert, de la part de l'utilisateur, une attention particulière pour suivre son itinéraire, ce qui le met dans des conditions plus proches de la marche dans un environnement réel. De plus, elle est mieux adaptée à l'évaluation des systèmes ubiquitaires qui imposent de fait un déplacement physique de l'utilisateur dans le système. Kjeldskov et Stage [Kjeldskov & Stage, 2004] ont comparé ces deux techniques de simulation avec une évaluation statique (utilisateur assis à une table) et une évaluation de terrain (utilisateur laissé libre de ses déplacements). Il en ressort des résultats a priori surprenants : toutes les techniques ont à peu près le même pouvoir de détection, sauf l'évaluation statique qui détecte nettement plus de problèmes d'utilisabilité que les autres. Ceci se manifeste notamment par un plus grand nombre de problèmes de niveau cosmétique[1] détectés. Les auteurs ont cependant identifié des biais possibles concernant le contexte expérimental, sur lesquels nous reviendrons.

Évaluation à distance

La mobilité de l'utilisateur, si elle n'est pas simulée par un appareillage, implique nécessairement un déplacement de celui-ci et donc l'augmentation (même transitoire) de la distance entre lui et les observateurs. Ainsi, l'évaluation d'un système mobile ou ubiquitaire relève aussi de problématiques très proches de l'évaluation à distance [Hammontree, Weiler, & Nayak, 1994]. Cette distance n'est pas forcément importante. Dans une rue, l'observateur peut être seulement à quelques mètres de l'utilisateur. Dans un laboratoire d'utilisabilité, l'utilisateur va se déplacer, se tourner, et donc rendre très difficile son suivi à travers une glace sans tain ou même avec des caméras orientables télécommandées. Dans tous les cas, la visualisation des interactions est délicate.

Lorsque la distance augmente, à partir de quelques dizaines de mètres, on ajoute à cela des difficultés pour collecter l'information. On a alors recours à des caméras de type col-de-cygne et des micros sans fil, systèmes comparables à ceux utilisés lors de certaines évaluations à distance [Scholtz, 2001]. Lorsque la distance augmente encore, dès que l'observateur perd le contact visuel direct avec l'utilisateur, la situation devient alors très similaire à une évaluation à distance, tant du point de vue technique que méthodologique. En effet, le rôle du facilitateur[2] de l'expérimentation s'en trouve modifié significativement car il doit utiliser des systèmes de communication (téléphone,

[1] La définition utilisée pour le terme « cosmétique » est celle de Molich [Molich, 2000].
[2] Le *facilitateur* est chargé d'accompagner l'utilisateur au cours de l'expérimentation, de lui donner des consignes, de le questionner, de résoudre d'éventuels petits problèmes techniques. Il se distingue de l'*observateur* qui a pour rôle unique d'observer la situation, de prendre des notes, de gérer les systèmes d'enregistrement, mais qui n'interagit pas directement avec l'utilisateur.

talkie-walkie, etc.) afin de transmettre les consignes et répondre aux questions de l'utilisateur.

Le domaine de l'évaluation ergonomique à distance (« Remote Usability Testing » pour les anglo-saxons) [Hammontree et al., 1994] est principalement représenté dans la littérature par des expérimentations destinées à l'évaluation de sites web. En effet, ces systèmes, assez simples à mettre en œuvre, sont bien adaptés à ce type de protocole expérimental. Nous n'avons trouvé dans la littérature qu'un seul exemple de test à distance d'un système mobile, mais là encore il s'agit d'un navigateur web sur PDA [Waterson, Landay, & Matthews, 2002]. La question centrale qui occupe bon nombre d'études sur l'évaluation à distance concerne la validité des analyses effectuées. Cette question est apparue très tôt [Hartson, Castillo, Kelso, & Neale, 1996] [Castillo, Hartson, & Hix, 1998] et continue d'alimenter la littérature avec des résultats contradictoires. Pour la majorité des auteurs, peu de différences sont constatées [Bartek & Cheatham, 2003] [Brush, Ames, & Davis, 2004] [Thompson, Rozanski, & Haake, 2004], pour d'autres au contraire, il y a des différences significatives sur les problèmes ergonomiques détectés [Tullis, Fleischman, McNulty, Cianchette, & Bergel, 2002]. De notre point de vue, ces contradictions sont issues du fait que, plus que la distance, ce sont les conditions expérimentales du terrain qui déterminent ces différences. Nous nous sommes alors posé la question de définir ce qu'est une expérience de « terrain ».

Définition de la notion de « terrain »

Dans la littérature, la notion de « terrain » n'est pas définie précisément, mais plutôt vue comme l'opposé du laboratoire d'utilisabilité. Il y a un consensus de fait à présenter les expérimentations de terrain comme revenant à placer l'utilisateur dans l'environnement « naturel » de l'usage attendu du système étudié. On parle parfois de situation « écologique ». Par exemple, un PDA destiné à des infirmières sera évalué à l'intérieur d'un hôpital [Kjeldskov, Skov, Als, & Høegh, 2004]. Dans le vocabulaire anglo-saxon, plusieurs termes sont utilisés pour décrire ce type de protocole expérimental. Les plus usités sont « field experiments » (e.g. [Goodman, Brewster, & Gray, 2004]), « field tests » (e.g. [Hertzum, 1999]), « field trials » (e.g. [Jensen & Larsen, 2007], ou expérimentations « in the wild » (e.g. [Waterson et al., 2002]). Certaines appellations plus anecdotiques sont parfois utilisées, comme « quasi-experimentations » [Roto, Oulasvirta, Haikarainen, Lehmuskallio, & Nyyssönen, 2004]. La communauté scientifique qui s'intéresse régulièrement à la question depuis quelques années, notamment via des ateliers, utilise également un vocabulaire varié. On parle ainsi de « reality testing »[3], d'expérimentations « in situ »[4] ou plus récemment de « mobile living labs »[5]. Il n'existe pas à notre connaissance de tentative d'uniformisation de ces appellations. Nous utiliserons le terme de « terrain » dans la première partie de l'article, puis nous le préciserons en introduisant la notion de « situation réelle », et enfin nous proposerons des appellations plus précises dans la partie discussion.

[3] CHI'2006 workshop « Reality Testing: HCI Challenges in Non-Traditional Environments » (http://www.cs.indiana.edu/surg/CHI2006/)
[4] MobileHCI'07 workshop « In-Situ 2007 : Using Mobile Devices and Emergent Technology for In-Situ Evaluations » (http://insitu2007.freeband.nl/)
[5] MobileHCI'09 workshop « Mobile Living Labs 09: Methods and Tools for Evaluation in the Wild » (http://mll09.telin.nl)

Quelle est la valeur ajoutée du terrain ?

De nombreux travaux ont cherché à déterminer si le fait de se placer dans le contexte réel apportait véritablement une plus grande validité aux évaluations [Fields, Amaldi, Wong, & Gill, 2007]. Parfois, il s'agit plus simplement de réduire les coûts logistiques de l'expérimentation [Rowley, 1994]. Nous avions précédemment réalisé un état de l'art de ces travaux [Jambon et al., 2006] et [Jambon et al., 2008]. Cependant, une recherche plus approfondie des publications existantes, de récents résultats issus de la littérature, ainsi que les retours d'expérience de nos propres expérimentations, nous ont amenés à remettre en cause certains points importants de cet état de l'art. Dans la littérature, la comparaison entre les protocoles expérimentaux s'effectue principalement selon deux critères : le nombre de problèmes détectés et éventuellement leur impact, généralement selon la définition de Molich [Molich, 2000]. Les résultats de ces travaux peuvent être classés en trois catégories : ceux qui ne détectent pas de différence notable, ceux qui estiment que le laboratoire apporte plus de détection de problèmes d'utilisabilité que le terrain, et ceux qui affirment le contraire.

Concernant la première catégorie, Kjeldskov et al. indiquent détecter très peu de différences entre le laboratoire et le terrain lors de l'évaluation d'un dispositif mobile destiné à des infirmières [Kjeldskov et al., 2004]. En effet, tous les problèmes d'utilisabilité, sauf un, ont été détectés en laboratoire. Une deuxième publication de Kjeldskov et al. concernant l'évaluation d'un guide mobile pour un réseau de transport en commun, confirme ce résultat quantitatif [Kjeldskov et al., 2005]. Cependant, les auteurs précisent qu'il y a peu de corrélation entre les problèmes détectés lorsqu'ils sont de niveau cosmétique. Les résultats quantitatifs sont similaires à ceux de Betiol et Cybis [Betiol & Cybis, 2005] ainsi que ceux de Kaikkonen el al. [Kaikkonen, Kekäläinen, Cankar, Kallio, & Kankainen, 2005]. Cependant, alors que Betiol et Cybis détectent une plus grande sévérité des problèmes détectés en laboratoire, Kaikkonen et al, au contraire, indiquent que les problèmes détectés sur le terrain sont légèrement plus sévères.

D'autres auteurs, parfois les mêmes que précédemment, affirment détecter plus de problèmes d'utilisabilité en laboratoire que sur le terrain. Par exemple, Kjeldskov et Stage ont réalisé une étude poussée concernant six méthodes : avec utilisateur assis à une table, en simulant la mobilité en laboratoire (4 façons différentes), et sur le terrain [Kjeldskov & Stage, 2004]. Ces travaux montrent que la technique la plus simple (utilisateur assis à une table) détecte plus de problèmes que toutes les autres, notamment le terrain. Remarquons cependant que le différentiel est principalement dû à des problèmes de niveau cosmétique. Ces résultats sont cohérents avec ceux de Baillie et Schatz [Baillie & Schatz, 2005]. Notons qu'ils sont aussi en accord avec les travaux de Hertzum sur l'évaluation à distance de système non mobiles [Hertzum, 1999].

Enfin, plus récemment, Duh et al. [Duh, Tan, & Chen, 2006] ont remis en cause ces études en affirmant qu'au contraire, le terrain apporte plus de détection de problèmes ergonomiques que le laboratoire d'utilisabilité. Le travail de Duh et al. est d'autant plus intéressant que les auteurs utilisent la même classification [Molich, 2000] de la sévérité des problèmes ergonomiques que Kjeldskov et al. [Kjeldskov & Stage, 2004] avec un protocole expérimental très similaire à celui de Kaikkonen et al. [Kaikkonen et al., 2005] qui concluaient au contraire que les différences étaient inverses ou minimes... Ces résultats sont cohérents avec les travaux de Po et al. concernant l'évaluation heuristique. Po et al. ont détecté que l'évaluation heuristique sur le terrain, c'est-à-dire lorsque les experts réalisent leur évaluation dans l'environnement d'usage prévu du dispositif,

permet de révéler de plus nombreux problèmes d'utilisabilité que lorsque cette évaluation est réalisée en laboratoire, notamment vis-à-vis des problèmes liés au contexte [Po, Howard, Vetere, & Skov, 2004].

Ces travaux mènent à la conclusion que non seulement il n'y a pas de consensus concernant l'intérêt des expérimentations sur le terrain, mais aussi que les résultats expérimentaux des comparaisons entre laboratoire et terrain sont difficilement reproductibles. Nous en avons conclu que les variables indépendantes permettant de différencier les expériences de terrain de celles en laboratoire ne sont pas encore bien identifiées ou que certains biais non identifiés faussent les résultats.

Biais possibles

En marge des résultats de leurs articles, les auteurs indiquent quelques remarques ou anecdotes qui peuvent donner des pistes intéressantes. Kjeldskov et al. [Kjeldskov & Stage, 2004] par exemple, indiquent qu'une explication possible du faible nombre de problèmes détectés sur le terrain peut venir du fait que les utilisateurs n'explorent pas tous les aspects de l'interface. En effet, ils n'effectuent que les tâches qui sont pertinentes vis-à-vis de leur activité sur le terrain.

Dans le même article, les auteurs ont remarqué que, lorsqu'un sujet se déplace dans une rue en compagnie d'un facilitateur et d'un caméraman, un effet de groupe se produit et l'utilisateur se retrouve isolé du monde, comme dans une bulle protectrice : les personnes de l'environnement s'écartent sur le passage et les relations sociales s'en trouvent limitées car les autres personnes n'osent pas s'immiscer dans l'expérimentation. Nous avons constaté le même effet lors de nos expérimentations, où les personnes connaissant un utilisateur ont eu des réticences à le saluer lors de son passage.

Baillie et Schatz ont détecté que les utilisateurs effectuaient leurs tâches plus rapidement et en faisant moins d'erreurs sur le terrain [Baillie & Schatz, 2005]. Ils supposent que les utilisateurs se sentaient plus « détendus » hors du contexte du laboratoire d'utilisabilité. De manière similaire, mais pour un système non mobile, Schulte-Mecklenbeck et al. [Schulte-Mecklenbeck & Huber, 2003] indiquent que, dans une tâche de recherche d'information via une interface web, les utilisateurs présents en laboratoire plutôt qu'à distance abandonnent moins facilement l'expérimentation et récupèrent plus d'informations. Les auteurs supposent que la « pression psychologique » de la présence de l'observateur en est la principale raison.

Une autre anecdote, rapportée par Isomursu et al. [Isomursu, Kuutti, & Väinämö, 2004] nous donne une autre piste. Dans son expérimentation, les utilisateurs sont recrutés dans la rue et par groupe de deux. L'un est chargé d'utiliser le dispositif (un PDA sensible au contexte) tandis que l'autre est chargé de filmer, à l'aide d'un téléphone mobile disposant d'une caméra, la situation vécue lorsqu'une difficulté intervient. Cette vidéo est nommée « experience clip ». Les auteurs ont constaté que lorsque les sujets étaient recrutés seuls, et que l'un des chercheurs servait de caméraman, la qualité, la durée et la pertinence des vidéos diminuait significativement. Plus surprenant, les sujets ont tous indiqué que cette présence ne les gênait en rien dans l'évaluation…

Limites des protocoles

Ces remarques et anecdotes nous ont amenés à étudier plus précisément les protocoles expérimentaux utilisés en laboratoire et sur le terrain. Dans toutes les études

citées concernant des systèmes mobiles, les protocoles sont quasiment identiques en laboratoire et sur le terrain. Notamment, les tâches effectuées par l'utilisateur sont prescrites. De plus, afin de faciliter le déroulement de l'expérimentation et l'observation, les évaluations se déroulent en présence d'un facilitateur, parfois aussi en présence d'un caméraman ou d'observateurs. L'instrumentation des expérimentations se base généralement sur l'usage d'une caméra filmant le contexte et parfois d'une seconde caméra col-de-cygne fixée au dispositif permettant de visualiser clairement les interactions entre l'utilisateur et le dispositif mobile.

L'intérêt d'utiliser des protocoles identiques est de pouvoir comparer les deux approches en ne considérant pas le protocole expérimental comme une variable indépendante. Cependant, du point de vue méthodologique, cela revient à utiliser le terrain comme une extension du laboratoire au sens où l'utilisateur, le dispositif, le facilitateur et l'instrumentation nécessaires à l'évaluation sont simplement « déplacés » dans un contexte réel. Seul le « décor » a changé. En conséquence, l'utilisateur se retrouve dans un contexte qui est loin d'être réel. En effet, l'utilisateur met en œuvre un dispositif qui est parfois très impressionnant et encombrant, notamment si une caméra col-de-cygne est utilisée. De plus, il subit, même sans en être conscient, une certaine influence due à la présence des personnes chargées de gérer ou d'observer l'expérimentation.

Au final, notre hypothèse est que sur le terrain, l'utilisateur évolue dans une sorte de « bulle » où les tâches qu'il doit réaliser, son activité générale, ses relations sociales, etc. sont contraintes par le protocole expérimental. Nous supposons que les bénéfices attendus du terrain sont en partie annihilés par ces contraintes, et qu'en conséquence, les différences détectées entre les deux configurations sont en partie des artefacts.

3. PROPOSITION D'UNE NOUVELLE APPROCHE

3.1. Laboratoire, terrain et situation réelle

Notre hypothèse initiale est qu'il n'y a pas deux configurations possibles mais plutôt trois : le laboratoire, le terrain et la **situation réelle**. Cette dernière configuration se différencie du terrain par une absence quasi totale des contraintes liées à l'expérimentation, permettant ainsi d'en limiter les artefacts. En effet, nous supposons que les différences entre le laboratoire et le terrain sont peu déterminantes car le protocole expérimental est finalement très similaire, du fait notamment de la présence de personnels et dispositifs d'observation, mais aussi du fait que la tâche est prescrite. Sur ces points, nos hypothèses sont en accord avec celles de Thomas et Kelloggs, exprimées dans les années 80, concernant la minimisation du « fossé écologique » [Thomas & Kellogg, 1989]. Ainsi, nous proposons de définir les approches « terrain » et « situation réelle » selon les critères suivants :

Comparée au laboratoire, la situation est dite **sur le terrain** lorsque :

- Le contexte d'usage du dispositif est réel, c'est-à-dire que l'évaluation doit avoir lieu dans le contexte où le dispositif est censé être utilisé, ou à défaut, un contexte équivalent ;
- Le dispositif et les données à disposition des utilisateurs sont perçus comme réels par l'utilisateur, même si des techniques de simulation comme celle du Magicien d'Oz sont utilisées.

La situation est dite **en situation réelle** si, de plus :

- La tâche de l'utilisateur n'est pas prescrite, mais découle de la situation vécue par l'utilisateur en fonction de ses propres aspirations et des évènements planifiés ou opportunistes issus du contexte ;
- Les personnels liés à la gestion de l'expérimentation et à l'observation, ainsi que les dispositifs d'acquisition sont hors du champ de perception de l'utilisateur.

De nombreuses expérimentations sur le terrain ont été réalisées grâce à des systèmes d'acquisition techniquement élaborés, notamment via des systèmes d'enregistrement audio-vidéo multi-sources portables (par exemple [Lyons & Starner, 2001], [Roto et al., 2004], [Zouinar, 2004] et [Calvet, Salembier, Kahn, & Zouinar, 2005]), grâce à des systèmes d'analyse de traces et de questionnaires proactifs (par exemple [Froehlich, Chen, Consolvo, Harrison, & Landay, 2007]) ou grâce à une combinaison de ces systèmes [Riegelsberger & Nakhimovsky, 2008]. Cependant, peu d'expérimentations en situation réelle sont relatées dans la littérature concernant les systèmes mobiles ou ubiquitaires. Les rares expérimentations concernant les téléphones mobiles se focalisent sur des statistiques d'usage comme celle de Demumieux et Losquin [Demumieux & Losquin, 2005] ou celle de Jensen et Larsen [Jensen & Larsen, 2007]. Citons également l'utilisation d'une technique dérivée du magicien d'Oz par Consolvo et al. pour la simulation in-situ d'un système ubiquitaire [Consolvo et al., 2007]. Cette rareté peut s'expliquer par le fait que ces expérimentations sont relativement difficiles à mettre en œuvre à cause de l'imprévisibilité des événements du contexte réel, un point souligné par Kellar et al. [Kellar et al., 2005], et aux difficultés rencontrées pour observer l'interaction sans introduire dans le contexte des systèmes d'acquisition visibles.

3.2. Principe d'incertitude et technique du cheval de Troie

Principe d'incertitude

En effet, en l'absence de facilitateur, d'observateur, de caméraman ou de dispositifs visibles d'acquisition de données, l'analyse des interactions entre l'utilisateur et le système devient difficile. On doit se limiter le plus souvent aux traces enregistrées par les dispositifs (actions de l'utilisateur sur l'interface, géolocalisation, capteurs, etc.) et aux entretiens postérieurs à l'expérimentation. Sont également exclus du protocole l'usage d'un journal de bord, la technique des incidents critiques [Hagen et al., 2005], ou d'autres techniques similaires comme celle des « emoticons » [Arhippainen, Rantakokko, & Tähti, 2004] car ils modifient l'activité de l'utilisateur et donc perturbent le caractère « écologique » de la situation.

Si, via les traces, la détection des principaux problèmes ergonomiques est possible, nous avons montré que leur explicitation reste un problème [Jambon et al., 2006]. Les entretiens qualitatifs à la fin de l'expérimentation sont bien adaptés à des expérimentations courtes où les objectifs sont relatifs à l'usage des dispositifs, mais ils sont beaucoup moins adaptés à des expérimentations de longue durée ou si l'on se focalise sur les détails de l'interaction, car l'utilisateur risque d'oublier une partie des difficultés rencontrées.

Nous nommons cette contrainte « Principe d'Incertitude » du fait de sa proximité sémantique avec celui énoncé par Heisenberg concernant la physique quantique, et qui dans notre contexte, peut s'énoncer ainsi : ***« il n'est pas possible à la fois d'observer précisément une situation d'interaction homme-machine sans, par effets de bord, la perturber »***. Un principe similaire peut se rencontrer en ethnologie [Malaurie, 2002]. Ce principe a pour conséquence de placer les expérimentateurs devant un difficile

dilemme : soit ils choisissent d'observer avec précision en acceptant de nombreux biais, soit ils minimisent les biais, mais ils ne disposeront alors que d'observations très limitées, avec pour conséquence de minimiser l'intérêt de l'étude.

Même s'il n'est pas a priori possible de transgresser ce principe d'incertitude, notre objectif a été de définir un protocole expérimental permettant de minimiser les biais tout en garantissant que nous disposerions d'assez d'informations pour appréhender la situation et ainsi réaliser l'évaluation. Pour des raisons déontologiques évidentes, il n'est bien entendu pas possible de faire abstraction du fait même que l'utilisateur ait connaissance d'être dans le cadre d'une expérimentation. Nous avons donc tout d'abord cherché à tirer le maximum des traces issues des dispositifs, puis nous nous sommes intéressés au moyen d'introduire des dispositifs d'observation sans qu'ils soient vus comme tels.

Technique du cheval de Troie

En situation réelle, il est ainsi nécessaire de se passer d'un observateur et a fortiori d'un caméraman. Pourtant, les enregistrements vidéos sont une source extrêmement riche et utile pour interpréter les événements non anticipés. Les caméras col-de-cygne, permettant de capturer l'interaction avec les dispositifs mobiles, sont un exemple de dispositif très utile mais trop intrusif. L'usage de caméras de très petite taille portées par l'utilisateur lui-même pourrait être une solution. Cette instrumentation reste cependant problématique car, aujourd'hui encore, ces caméras ne sont pas assez performantes et miniaturisées pour se faire oublier. De plus, la présence de la caméra doit être indiquée à l'utilisateur pour d'évidentes raisons déontologiques.

C'est pourquoi, plutôt que de chercher à masquer les dispositifs d'instrumentation, notre idée a été de les faire passer pour autre chose que ce qu'ils sont. Cette technique est de la même inspiration que celle du magicien d'Oz, au sens où l'on masque certains aspects du protocole d'évaluation à l'utilisateur. Nous l'avons nommée « Cheval de Troie », car elle est basée sur un double usage des dispositifs liés à l'instrumentation : le premier usage est voyant, mais sa présence est justifiée, le second et véritable usage est masqué. Ainsi, pour un dispositif d'instrumentation donné, nous présentons à l'utilisateur ce dispositif comme faisant partie du système ou de l'environnement étudié. Mais en fait, l'intérêt principal du dispositif est l'acquisition des informations nécessaires à l'évaluation.

Par exemple, pour l'évaluation d'un dispositif destiné à des skieurs [Jambon, 2006], la mini-caméra disposée sur le casque de chacun des skieurs avait pour premier usage d'enregistrer le vécu du skieur afin de pouvoir le lui rejouer sur son smartphone. Cette mini-caméra permettait aussi de filmer l'interaction du skieur avec son smartphone, et de ce fait, elle permettait de disposer d'informations très utiles pour l'évaluation. Cette astuce nous avait semblé être une solution ponctuelle. Nous l'avons néanmoins reproduite, sous une forme différente, lors de l'expérimentation au Muséum de Lyon : la carte distribuée aux visiteurs faisait partie de la scénographie et permettait, en même temps, l'étude du parcours des visiteurs [Jambon, Mandran et al., 2007].

Nous nous sommes ainsi aperçus que la technique du cheval de Troie était généralisable dans de nombreux cas. Deux approches sont possibles :

- Si le système ou l'environnement dispose déjà d'un dispositif ayant potentiellement les capacités d'acquisition des données souhaitées, il s'agit de le compléter, par exemple par un système d'enregistrement ou par l'installation

d'équipements complémentaires. Dans ce cas, la justification de la présence du dispositif est aisée.

- Si le système ou l'environnement ne contient pas ce dispositif, il faut l'y ajouter en le faisant passer pour une fonctionnalité supplémentaire, mise à disposition de l'utilisateur, même si elle ne sert à rien du point de vue de l'expérimentation. Par exemple, si l'on souhaite étudier l'usage géolocalisé des fonctions d'un téléphone mobile, il faudra que le téléphone dispose aussi d'un GPS mis à disposition de l'utilisateur et présenté comme l'une des fonctionnalités testées, même si l'expérimentation ne s'intéresse qu'à l'envoi de SMS.

Il existe néanmoins certains cas où cette technique n'est pas utilisable. De manière générale, elle ne peut être mise en œuvre dans les cas où la présence du dispositif d'acquisition n'est pas justifiable ou semble incongrue. Par exemple, il sera délicat de justifier l'usage d'une mini-caméra portée par l'utilisateur si l'on souhaite évaluer un dispositif grand public destiné à donner les horaires de bus.

La technique du cheval de Troie pose également un sérieux problème déontologique. En effet, si les informations données à l'utilisateur ne sont pas complètes, cette technique revient à enregistrer l'activité de l'utilisateur à son insu, ce qui est n'est pas en accord avec les règles d'éthique des expérimentations [Johnson, Solso, & Beale, 1997]. Ainsi, une attention particulière doit être portée à l'aspect déontologique du protocole expérimental. Il est notamment impératif d'informer les utilisateurs de l'ensemble des données recueillies et enregistrées, et de l'utilisation qui va en être faite. Cela est délicat, car afin de maintenir l'aspect « écologique » de la situation, il serait préférable de cacher aux utilisateurs le véritable objectif de l'acquisition de données. Sur le plan déontologique, ce n'est pas admissible sauf dans certains cas très particuliers d'expérimentations en psychologie expérimentale. C'est pourquoi nous proposons une voie médiane, où nous informons les utilisateurs de l'utilisation des données en « remarques » lors des consignes. L'information complète leur est donnée en fin d'expérimentation, moment auquel leur consentement pour l'utilisation réelle de ces données peut être recueilli. Notons que si des données personnelles sont nécessaires, une déclaration à la CNIL doit être effectuée.

3.3. Motivations et agenda de recherches

Notre hypothèse est qu'il existe trois configurations pour l'évaluation des dispositifs mobiles. Si la première, le laboratoire, est aujourd'hui bien connue et validée, les deux autres restent encore à défricher aussi bien du point de vue technique que méthodologique. De plus, il n'y a pas unanimité concernant le gain attendu à utiliser une configuration sur le terrain ou en situation réelle vis-à-vis d'une configuration en laboratoire. Les configurations laboratoire et terrain ayant déjà été bien étudiées dans la littérature, nous nous sommes intéressés plus particulièrement à la situation réelle. Nous souhaitons en effet déterminer si ces expérimentations en situation réelle peuvent apporter certaines informations impossibles à obtenir en laboratoire ou sur le terrain. Notre étude s'intéresse aux configurations expérimentales, et plus particulièrement à l'aspect recueil. Nous ne nous sommes pas intéressés aux autres caractéristiques du protocole comme la sélection des sujets ou les techniques d'évaluation.

Nous nous sommes ainsi engagés dans une série de trois expérimentations destinées à valider les difficultés techniques et méthodologiques des expérimentations en situation réelle. Une des difficultés de cette démarche est l'impossibilité de reproduire, dans les

trois configurations, le même protocole expérimental. En effet, en situation réelle, les systèmes d'acquisition classiques utilisés en laboratoire sont proscrits. C'est pour cette raison que, dans un premier temps, notre objectif a été d'évaluer le pouvoir d'analyse des traces que peuvent laisser les dispositifs comparés aux techniques d'observation classiques (enregistrements audio-vidéo principalement). Pour cela, nous avons réalisé une première expérimentation faisant appel à deux techniques d'évaluation distinctes : MapMobile [Jambon et al., 2006]. Dans un second temps, nous avons cherché à valider les protocoles expérimentaux concernant la configuration en situation réelle peu explorée dans la littérature. Pour cela, deux expérimentations, dont le but était de tester la capture de données en situation difficile (activité sportive en extérieur) ou sur une longue durée (quelques semaines) et un grand nombre d'utilisateurs (plusieurs centaines de personnes), ont été mises en place : E-Skiing [Jambon, 2006] et Muséum [Jambon, Mandran et al., 2007].

4. EXPERIMENTATIONS ET RETOURS D'EXPERIENCE

Les trois expérimentations décrites ci-après ont été effectuées dans le cadre des projets ADAMOS (RNTL-PROACT Franco-Finlandais) et IMERA (Région Rhône-Alpes Emergence). Ces deux projets avaient pour objectif l'étude du processus de conception de services pouvant réagir de manière proactive vis-à-vis des utilisateurs. Ces services s'appuient sur des dispositifs mobiles, ubiquitaires, communicants, et sensibles au contexte d'usage, permettant ainsi aux services d'être adaptatifs. Au cours de ces projets, trois campagnes d'expérimentation des dispositifs ont été effectuées.

4.1. Expérimentation « MapMobile »

MapMobile est un assistant numérique personnel permettant à son utilisateur de se géolocaliser et d'être guidé à l'intérieur et à l'extérieur d'un bâtiment. Le prototype a été développé conjointement par France Télécom R&D et le CEA-Leti. Dans le cadre du projet ADAMOS, des fonctions d'adaptativité (en fonction des centres d'intérêts de l'utilisateur) et de proactivité (en fonction de la géolocalisation et d'évènements extérieurs) ont été ajoutées. L'objectif de l'expérimentation était triple : (1) évaluer l'utilisabilité du prototype, (2) valider les méthodes et techniques d'évaluation et (3) servir d'illustration à une étude sociologique. Ce sont seulement les deux premiers points qui nous intéressent ici.

Protocole expérimental

L'expérimentation était située dans un environnement réel : les bâtiments de France Télécom R&D à Meylan pendant les heures de travail. Le dispositif était totalement fonctionnel. Une tâche de haut niveau était imposée à l'utilisateur : rendre visite à un contact professionnel dans un lieu inconnu de lui. Au cours de l'expérimentation, l'utilisateur était accompagné d'un facilitateur et d'un caméraman (cf. images 1 et 2). Il s'agissait donc d'une expérience de terrain sans être une expérimentation en situation réelle. En effet, nous avons cherché des conditions expérimentales permettant de disposer à la fois des données classiques disponibles lors d'évaluations sur le terrain et une simulation des données disponibles lors d'évaluations en situation réelle.

Images 1 et 2 : sujet (au centre en bleu) avec le facilitateur (à droite en blanc) et le caméraman (à gauche en orange) filmant l'écran du système MapMobile (image de droite) au cours de l'expérimentation.

L'expérimentation a concerné une douzaine de sujets dont seulement dix ont été effectivement inclus dans les analyses du fait de problèmes techniques ou organisationnels. Les sujets ont été sélectionnés de manière à être répartis selon les critères sociologiques de la méthode CAUTIC [Forest, Guilloux, Mallein, & Panisset, 1998]. Dans les faits, même si cela n'était pas recherché, l'expérience des sujets dans l'usage des nouvelles technologies ainsi que leur genre ont été également bien répartis. Les sujets étaient rémunérés. Les bâtiments de France Télécom Meylan R&D ont la forme de carrés, vides en leur centre, se joignant par les sommets sans marques distinctives. L'orientation à l'intérieur des bâtiments est particulièrement difficile, ce qui rendait le contexte d'expérimentation tout à fait réaliste[6].

Le protocole était le suivant : Les sujets étaient accueillis au poste de garde situé à l'entrée du site, puis accompagnés jusqu'à l'accueil du bâtiment. À ce moment, le PDA destiné à les guider dans le bâtiment leur était remis et ils étaient équipés d'un micro-cravate sans fil. Les consignes leur étaient alors données. Ensuite, les sujets se déplaçaient librement dans le bâtiment, guidés par le dispositif, tout en étant régulièrement interrompus par celui-ci du fait de son comportement proactif. Il était notamment fait mention d'un retard de la personne visitée, d'un document à aller chercher, d'informations profilées disposées sur un panneau d'affichage, etc. L'expérimentation durait environ une demi-heure. Tout au long de l'expérimentation, le facilitateur encourageait les sujets à exprimer leurs difficultés et ils étaient filmés par un caméraman. À l'issue de l'expérimentation, les sujets étaient reçus en entretien par un sociologue.

Évaluation et méta-évaluation

L'objectif de l'expérimentation pour notre équipe était donc double : évaluer l'utilisabilité du dispositif et comparer deux méthodes d'évaluation selon les types de données disponibles. Nous avons classé ces données disponibles selon deux types. Les traces dites « terrain » se réfèrent aux données habituellement prises en compte sur le terrain, c'est-à-dire les enregistrements audio-vidéo du sujet et du dispositif ainsi que le film de son écran. Ces données sont caractérisées par leur grande richesse, mais aussi par le fait qu'il est très difficile de les analyser automatiquement. Au contraire, les traces

[6] Pour l'anecdote, les expérimentateurs se sont parfois eux aussi perdus dans les couloirs…

dites « situation réelle » regroupent tous les événements issus de l'interaction entre l'utilisateur et le système (appuis sur les boutons, sélections de commandes, affichages, sons). Ces données, peu riches et le plus souvent de faible niveau d'abstraction, sont peu lisibles manuellement mais plus aisées à analyser automatiquement.

Nous avons cherché à déterminer le pouvoir d'analyse de chaque type de trace, mais aussi et surtout à déterminer si les problèmes ergonomiques découverts à l'aide des traces « situation réelle » peuvent être confirmés par les traces « terrain » dont la méthodologie est bien connue. Notre objectif était de déterminer s'il était possible de s'abstraire des traces « terrain », difficiles à obtenir en situation réelle, sans pour autant perdre la validité des résultats obtenus. Pour cela, nous avons suivi une démarche en deux phases :

1. **Analyses en aveugle :** Nous avons analysé les traces « terrain » (film d'écran, notes et fichiers audio-vidéo) et les traces « situation réelle » (actions de l'utilisateur sur le système et retours d'information du système) de manière indépendante sans avoir connaissance des informations issues de l'autre type de traces. L'objectif était de déterminer ce que l'on peut déduire de chacun des deux types de traces pris isolément, c'est-à-dire leur réel pouvoir d'analyse.
2. **Croisement des résultats :** À partir des conclusions des deux analyses précédentes, nous avons croisé les résultats obtenus afin de détecter les cohérences, les incohérences, et les résultats complémentaires. Nous avons ensuite catégorisé les résultats selon les fonctions étudiées, les critères ergonomiques analysés, etc.

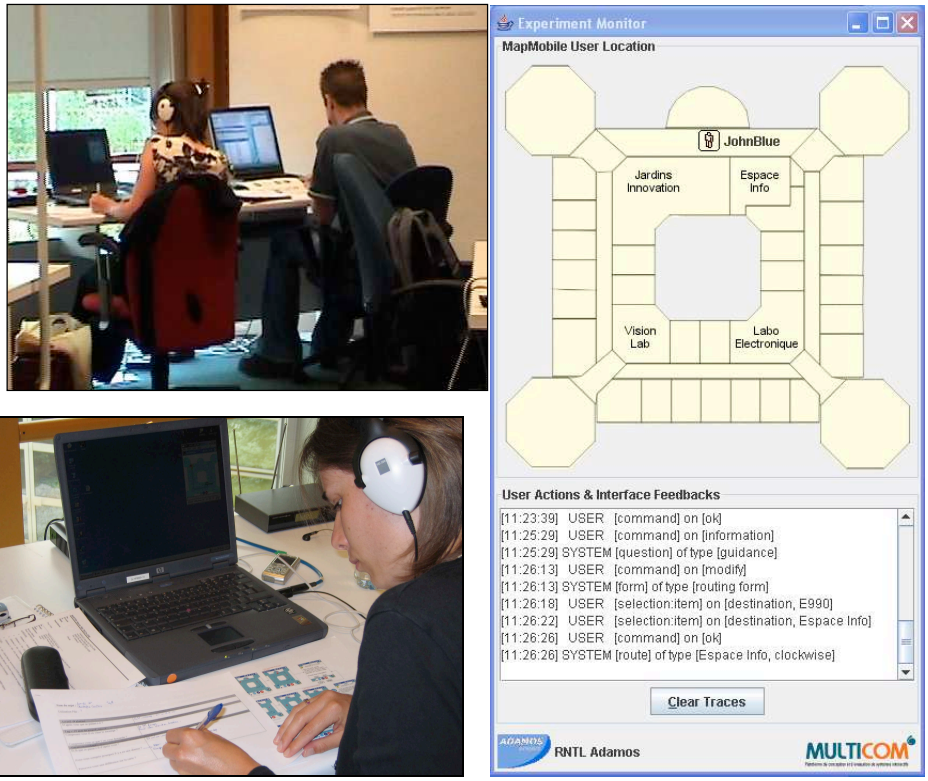
En pratique, les deux types de traces étaient recueillis tout au long de l'expérimentation (à l'exception des images du caméscope) dans une régie improvisée dans un bureau du bâtiment, via des technologies sans fil (HF et WiFi). Cette régie n'était pas montrée aux sujets de l'expérimentation (cf. images 3, 4 et 5).

Évaluations « terrain » et « situation réelle »

Aucune consigne particulière n'avait été donnée aux évaluateurs concernant le fond ou la forme des résultats attendus. Il leur était simplement demandé de détecter le plus de problèmes d'utilisabilité possible, en se limitant impérativement à l'usage d'un seul des deux sous-ensembles de données disponibles. Les résultats dits « terrain » ont été obtenus à partir des données suivantes :

- Les enregistrements audio et vidéo du contexte réalisés par le caméraman à l'aide d'un caméscope grand public ;
- L'enregistrement audio des commentaires du sujet obtenus via un micro-cravate sans fil et enregistrés sur un ordinateur portable situé dans la régie ;
- L'enregistrement du film d'écran du PDA utilisé par le sujet, obtenu via une liaison réseau sans fil et enregistré sur un ordinateur portable situé dans la régie.

Les commentaires du sujet étaient écoutés (via un casque) et annotés au fil de l'eau par une ergonome qui disposait en plus du film d'écran du PDA en temps réel. Cette ergonome se trouvait dans la régie et donc hors de la vue du sujet. Un rapport a ensuite été rédigé à partir de ces notes en se basant notamment sur les critères ergonomiques classiques [Bastien & Scapin, 1995].



Images 3, 4 et 5 : sur l'image en haut à gauche, régie permettant le recueil des traces « terrain » (à gauche) et « situation réelle » simulée (à droite). Sur l'image en bas à gauche, ergonome effectuant les annotations au fil de l'eau. Sur l'image de droite, copie d'écran des traces « situation réelle » recueillies pendant l'expérimentation.

Les résultats dits « situation réelle » ont été obtenus à partir des traces informatiques collectées sur le dispositif. Les traces étaient transmises en temps réel par le réseau sans fil via un bus de données développé spécifiquement (Usybus basé sur Ivy [Buisson et al., 2002]). Les traces collectées étaient :

- Les actions de l'utilisateur sur l'interface du dispositif ;
- Les réactions de l'interface du dispositif perceptibles par l'utilisateur ;
- La géolocalisation du dispositif par triangulation WiFi.

Un ingénieur était chargé de surveiller le bon déroulement de l'acquisition et de l'enregistrement des données. Ces traces ont été stockées puis analysées a posteriori, principalement en construisant des matrices de transition et en utilisant des algorithmes de détection de différences en les comparant à une expérimentation de référence (sans erreur) effectuée par un membre de l'équipe.

Évaluation croisée « terrain » versus « situation réelle »

De manière globale, nous avons constaté, en première lecture, que le nombre de problèmes ergonomiques détectés grâce aux traces « terrain » est bien plus important que le nombre de problèmes détectés par l'analyse des traces « situation réelle ».

Les problèmes détectés à partir des traces « terrain » sont de nature absolue au sens où ces problèmes se réfèrent aux critères ergonomiques ou heuristiques connus. À

l'inverse, les traces « situation réelle » détectent des problèmes le plus souvent par comparaison entre différentes passations ou par rapport à une passation de référence. On peut donc les considérer comme plus relatives.

Le niveau d'abstraction des problèmes détectés par les traces « situation réelle » peut sembler au premier regard d'un faible niveau (cosmétique). En fait, ces traces permettent de détecter des anomalies au niveau des interactions, mais les problèmes détectés sont souvent de plus haut niveau (par exemple : des concepts mal compris). Ce faible niveau d'abstraction est aussi lié au fait que les traces enregistrées ne possédaient pas d'informations sur les tâches en cours (limite connue des traces « situation réelle »).

Les deux méthodes ont clairement montré que les utilisateurs n'ont pas compris le principe de devoir indiquer au système la présence d'un obstacle sur le chemin prévu (sur le terrain grâce aux commentaires / en situation réelle grâce à la détection d'une durée d'accomplissement de la tâche trop importante). De même, l'utilisation d'un point de passage lors de la spécification du routage est peu utilisée par les sujets (sur le terrain grâce aux commentaires / en situation réelle grâce à la détection d'actions non réalisées). L'incompréhension par les utilisateurs du concept de message géolocalisé a été clairement mis en évidence par les deux méthodes (sur le terrain grâce aux commentaires / en situation réelle grâce à la détection d'actions non réalisées).

Les traces « terrain » ont permis, principalement grâce aux commentaires des utilisateurs, de détecter de nombreux problèmes difficiles à identifier par d'autres moyens, par exemple : la signification parfois ambiguë des icônes, l'identification incorrecte de la modification du routage, l'orientation de la carte, la compréhension des messages affichés, etc. De manière générale, ces problèmes ont trait à l'interprétation de la part des sujets d'informations qui, même si elles ne sont pas correctement interprétées, ne provoquent pas de modification notable de la séquence d'interaction. Ceci est en partie lié au biais introduit dans le protocole par le rôle du facilitateur, qui, une fois le problème exprimé par le sujet, aidait celui-ci à poursuivre son interaction de manière correcte. Ainsi, c'était uniquement cette interaction correcte qui était enregistrée dans les traces « situation réelle », masquant de ce fait le problème rencontré par le sujet.

Un faux négatif a été découvert sur l'analyse de l'utilisation du guidage. En effet, trois des utilisateurs ont confondu la commande « modifier guidage » avec la commande « reprendre guidage ». Cette variation a été détectée, mais finalement jugée normale dans le processus naturel d'exploration du logiciel. Or, la confusion était bien réelle. Cette interprétation trop optimiste a été favorisée par le grand nombre de variations inter-utilisateurs détectées dans l'utilisation de la page concernée.

À l'opposé, trois types de faux positifs ont été mis à jour. Le premier type concerne la détection d'un temps anormalement long lors de l'indication d'un obstacle sur le parcours. En effet, un sujet a mis quasiment trois minutes pour acquitter le message d'avertissement. Un problème d'utilisabilité sérieux a donc été soupçonné. En fait, le message est arrivé un peu trop tôt dans le scénario tandis que le sujet et le facilitateur étaient engagés dans une discussion à propos de la précédente étape du scénario. Ici, le biais vient de la conjonction d'une limitation du prototype et du rôle du facilitateur. Le deuxième type de faux positif a été provoqué par la seule limite du prototype. En effet, certaines actions affichées n'étaient pas suivies des actions attendues, par exemple pour la commande « annuler ». Ainsi, nous avons détecté des appuis répétés sur cette commande, ce qui nous a fait soupçonner un problème de libellé mal compris ou de lenteur de réaction du dispositif. En fait, l'utilisateur pensant qu'il n'avait pas appuyé la première fois, a répété son action. Le troisième type de faux positif concerne la

détection d'actions manquantes. On a supposé, dans ce cas, que l'utilisateur n'avait pas suivi la trajectoire d'interaction prévue. En fait, un problème de transmission sur le réseau en était responsable.

Synthèse

En résumé, les traces obtenues en situation réelle ne permettent pas de détecter autant de problèmes d'utilisabilité que les traces obtenues sur le terrain. Ceci est particulièrement vrai en ce qui concerne les difficultés d'interprétation. Cette affirmation est à modérer du fait des biais introduits par le protocole (rôle du facilitateur et scénario très linéaire). Cependant, on peut se douter que les traces obtenues en situation réelle, même si elles détectent un nombre significatif de problèmes, ne permettront pas d'expliciter de manière fiable la cause de ces problèmes. En effet, le manque d'informations sur le contexte, que l'on obtient généralement grâce aux enregistrements audio-vidéo, reste un problème délicat qui peut engendrer à la fois des faux négatifs et des faux positifs.

Or, se limiter aux traces obtenues en situation réelle est très intéressant dans les situations où l'enregistrement de vidéos du contexte est difficile ou même non souhaitable. C'est pourquoi les informations de contexte qui font défaut doivent être reconstituées par d'autres moyens. Pour cela, nous proposons de récupérer le maximum de données du contexte par des capteurs additionnels (centrale d'attitude, géolocalisation, état du noyau fonctionnel de l'application, film d'écran) de manière à combler ce manque informationnel et ainsi reconstituer des bribes de contexte.

4.2. Expérimentation « E-skiing »

E-skiing est un service mobile destiné à enrichir la pratique du ski via l'envoi proactif aux skieurs, sur le terrain et peu de temps après leurs descentes, de données concernant celles-ci. Ces données permettent aux skieurs de revivre leurs précédentes descentes, via une vidéo et un ensemble de grandeurs numériques liées à leurs performances (trajectoire, vitesse, accélération, etc.). Pendant l'expérimentation, les informations issues des capteurs portés par le skieur étaient enregistrées pour chaque descente, traitées en fin de descente, puis les skieurs, de retour sur les pistes, étaient prévenus de manière proactive que de nouvelles informations sur leurs précédentes descentes étaient disponibles. Ils pouvaient alors consulter individuellement, grâce au smartphone (cf. images 6 à 10) : le chemin parcouru, la distance et le temps de descente, la vitesse maximale, le coefficient d'engagement (lié à l'accélération) et la vidéo de leur descente.

L'objectif de cette deuxième expérimentation était triple : (1) tester l'ergonomie de l'interface de restitution, (2) tester l'usage fait par un groupe de skieurs d'un service mobile proactif et (3) valider les aspects techniques et méthodologiques d'une évaluation en situation réelle.

Protocole expérimental

Pour cette expérimentation, nous avions recruté un groupe de skieurs ayant l'habitude de pratiquer ensemble, ceci afin de permettre également une approche sociologique des interactions dans le groupe. Ces skieurs étaient habitués à l'usage des nouvelles technologies et n'ont pas eu de difficulté à utiliser le matériel. Après une présentation des capteurs et des possibilités du service par une démonstration, les sujets

étaient invités à aller skier comme ils en avaient l'habitude, puis à revenir au local en bas des pistes après chaque descente (pour le chargement des données). Il s'agissait d'une expérimentation en situation réelle car les skieurs étaient libres d'effectuer les activités qu'ils désiraient et n'étaient pas accompagnés d'observateurs. L'expérimentation a mobilisé cinq skieurs dont quatre étaient équipés du service. Dans une seconde partie de l'expérimentation, qui n'est pas l'objet de cet article, les skieurs ont été conviés à une réunion sous forme de soirée conviviale de type « focus-group » où la version sédentaire du service leur était présentée, celle qui est destinée à être utilisée à la maison. Cette réunion avait également pour objectif d'obtenir un retour sur les aspects ergonomiques du système et sur l'usage qu'ils avaient fait du service proposé.

Images 6, 7 et 8 : Casque équipé d'une mini-caméra et d'un accéléromètre (à gauche), téléphone mobile implémentant l'interface de restitution mobile (à droite) et groupe de skieurs équipés (au centre).

Images 9 et 10 : Écran du téléphone mobile implémentant l'interface de restitution (à gauche) et copie d'écran de cette interface (à droite).

Techniquement, le service E-skiing était composé d'un ensemble de capteurs associés à des enregistreurs autonomes (vidéo, accélération, et position géographique) portés par les skieurs et d'une interface de restitution (téléphone mobile de type « smartphone ») qu'ils emmenaient aussi avec eux (cf. images 6 à 10). En back-office, nous disposions de deux ordinateurs portables destinés à traiter les données et d'un serveur de données permettant de mettre à disposition les données une fois traitées. La récupération des données s'effectuait à la fin de chacune des descentes car les enregistreurs autonomes n'avaient pas la capacité de se connecter sans fil. Par contre,

les messages proactifs et l'accès aux performances par les skieurs s'effectuaient via le réseau mobile (sauf les vidéos pour une raison de coût). Les détails des aspects techniques de l'expérimentation peuvent être consultés dans [Jambon, 2006] et [Jambon & Meillon, 2009].

Nous n'avons testé qu'un seul groupe de skieurs car l'objectif principal n'était pas d'obtenir des résultats statistiquement pertinents, mais s'approchait plus du test d'un démonstrateur de concept. Nous avons mis en œuvre la technique du « cheval de Troie » pour les caméras et les GPS portés par les skieurs. En effet, chacune des caméras permettait à la fois de filmer la descente des skieurs et leurs interactions avec les téléphones mobiles. Chacun des GPS permettait de calculer le parcours, mais aussi de localiser les lieux d'usage du service.

Principaux résultats

Le premier résultat de cette expérimentation est d'ordre technique. Cette expérimentation mettait en œuvre en situation réelle des technologies très innovantes et donc parfois peu fiables… Nous avions anticipé les problèmes techniques en effectuant de nombreux tests en laboratoire, en extérieur, puis sur les lieux. Malgré ces tests, les enregistrements vidéos se sont révélés quasiment inexploitables pour l'analyse des interactions du fait d'un mauvais cadrage : le faible angle de vue des mini-caméras et les déplacements des casques au cours des sauts effectués par les skieurs ont eu pour conséquence de placer hors champ les écrans des smartphones. Cet aspect n'avait pas été détecté lors des tests, car ceux-ci s'effectuant selon des tâches prescrites, les skieurs n'avaient pas fait de sauts comme ils en avaient pourtant l'habitude…

C'est pourquoi l'utilisabilité du service n'a pas pu être étudiée par ce moyen. Elle a été évaluée dans la seconde partie de l'expérimentation via des questions informelles posées au cours du focus-group. Nous avons été très surpris de constater, que, malgré une interface homme-machine développée très rapidement et sans attention particulière à sa qualité ergonomique, les utilisateurs n'ont fait aucune critique. Or, de manière évidente, l'ergonomie était largement perfectible. C'est en soi un résultat intéressant : les utilisateurs se sont focalisés sur le service et ont occulté les aspects ergonomiques qu'ils considéraient comme secondaires. Ainsi, il faut garder à l'esprit qu'une évaluation ergonomique réalisée en situation réelle, via des questionnaires, peut conduire à minimiser les problèmes d'utilisabilité réels du dispositif.

Malgré l'absence d'enregistrements vidéo utilisables, la géolocalisation et les logs des envois et réceptions de données nous ont permis d'obtenir des informations très intéressantes concernant l'usage du service proposé. L'activité des skieurs a été reconstruite à partir de ces informations qui avaient été enregistrées en premier lieu pour faciliter la mise au point du système. Nous avons ainsi pu déterminer les moments où les skieurs étaient prévenus par les relevés de SMS et où ils accédaient à l'information via les logs du serveur de données. Les enregistrements GPS nous ont ensuite permis d'en déduire les lieux. Nous nous sommes ainsi aperçus du très faible usage du dispositif en général, et de son utilisation quasi exclusivement hors des pistes. Ce résultat a priori surprenant a été explicité lors du focus group. Les skieurs étant venus pour skier, c'est effectivement ce qu'ils ont fait ! Ensuite, lors de pauses, ils ont parfois pris le temps de regarder si le service pouvait être intéressant. Ils ont également détourné l'usage de l'enregistrement vidéo, en se filmant entre eux au lieu d'utiliser la caméra pour filmer leur propre descente. Ces résultats mettent en lumière l'intérêt de l'approche en situation

réelle : les utilisateurs vivent leur vie et n'utilisent le système à leur disposition que s'ils en sentent le besoin. Nous sommes très loin des évaluations sur le terrain.

Les difficultés les plus importantes que nous avons rencontrées ont été d'ordre logistique et méthodologique. Le traitement des données en cours d'expérimentation a transitoirement saturé l'équipe d'observateurs stationnée au chalet. En effet, nous n'avions pas anticipé le fait que les skieurs reviendraient très régulièrement et ensemble au chalet. L'important volume de données à traiter rapidement, ajouté au fait que les enregistreurs ont parfois été mélangés entre les skieurs, a fait régner une certaine confusion. En outre, le protocole imposait la mise en route des enregistreurs par les skieurs eux-mêmes. Or, les consignes étant réduites, tous les skieurs n'ont pas toujours compris la procédure et ont parfois oublié les mises en route. De plus, ils n'ont pas osé utiliser les téléphones mobiles pendant certaines phases d'acquisition de données, pensant que ceux-ci n'étaient pas disponibles. Ce dernier point a été identifié comme un biais, mais les skieurs nous ont indiqué que cela n'avait eu aucune conséquence sur leur usage du service.

Synthèse

Ces expérimentations ont montré que la mise en œuvre d'une évaluation en situation réelle est véritablement très complexe aussi bien sur le plan technique que méthodologique. Nous avons appris à cette occasion qu'il est impératif de ne faire aucune concession sur ces deux aspects. En d'autres termes, le dispositif testé, comme les systèmes d'acquisition, doivent fonctionner de manière fiable, robuste, être autonomes et avoir une gestion automatique. Il ne faut pas déléguer une partie, même faible, du protocole aux utilisateurs. Ceci est une contrainte très forte, car les systèmes aujourd'hui disponibles pour les tests sont souvent des prototypes peu robustes et peu intégrés.

Nous avons également montré que la situation réelle est peu propice à la détection des problèmes d'utilisabilité, mais que de très intéressants résultats concernant l'usage réel des systèmes peuvent être aisément obtenus à partir des traces. Les résultats détaillés de cette expérimentation peuvent être consultés dans [Jambon, 2006]. Suite à cette expérimentation, nous avons cherché à valider un protocole expérimental adapté à l'évaluation en situation réelle d'un système ubiquitaire sur une longue durée et avec un grand nombre d'utilisateurs, afin de pouvoir disposer de données statistiquement intéressantes.

4.3. Expérimentation « Muséum »

L'expérimentation « Muséum » s'est déroulée pendant l'exposition temporaire « ni vu – ni connu »[7] au Muséum du département du Rhône. Cette exposition, dont le thème était le camouflage, se proposait d'enrichir le parcours du visiteur d'une expérience sur les concepts de vie publique et privée vis-à-vis de paparazzi virtuels. Au cours de son parcours dans l'exposition, le visiteur était amené à fournir des informations personnelles et à être photographié. À la fin de l'exposition, il lui était présenté la première page d'un journal à scandales fictif fusionnant les informations recueillies à son insu (cf. figure 1). Le visiteur était ensuite informé des techniques utilisées pour le « traquer » et des problématiques liées au respect de la vie privée.

[7] http://www.museum-lyon.org/expo_temporaires/ni_vu_ni_connu/

La demande d'évaluation de l'exposition était issue en premier lieu du scénographe de l'exposition, lequel désirait connaître l'affluence et plus précisément l'utilisation effective des bornes interactives ainsi que la fiabilité des techniques mises en œuvre pour implémenter la scénarisation. Plus généralement, les responsables des musées se préoccupent de connaître le comportement de leurs visiteurs selon les expositions proposées [Gob & Drouguet, 2006] et désirent également savoir comment les dispositifs interactifs installés sont utilisés. Or ces informations, pourtant basiques, sont très rarement disponibles et fiables. En effet, même si cela semble a priori simple, il n'est pas aisé d'analyser le parcours de visiteurs, car il ne suffit pas d'effectuer des comptages aux points de passage, il faut également pouvoir distinguer les visiteurs les uns des autres.

Notre objectif, par rapport à la précédente expérimentation E-skiing, était de prouver la possibilité de passage à l'échelle des expérimentations en situation réelle. Nous avons donc augmenté la durée de l'expérimentation : quelques semaines au lieu d'une seule journée pour les skieurs. Nous avons aussi augmenté le nombre d'utilisateurs : plusieurs centaines de visiteurs au lieu d'un seul groupe de quatre skieurs. Cela nous a permis d'obtenir des résultats statistiquement exploitables.

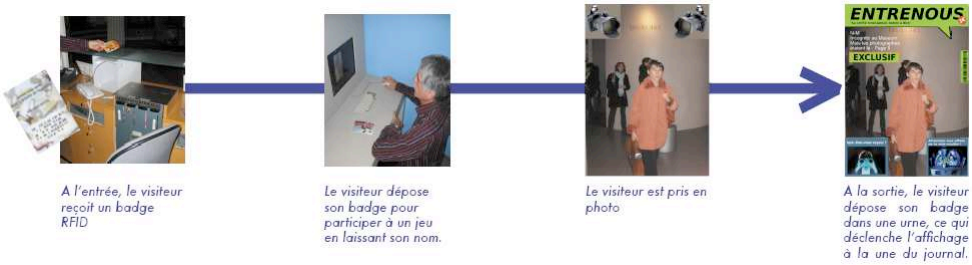


Figure 1 : Étapes du traçage des visiteurs dans l'exposition « ni vu – ni connu ».

Protocole expérimental

Le fonctionnement de cet environnement ubiquitaire se basait sur la technologie des étiquettes radiofréquence (RFID). Lors de son entrée dans le musée, le visiteur se voyait remettre, en même temps que son ticket d'entrée, une carte avec un texte à trous. Une étiquette RFID était dissimulée dans cette carte et des lecteurs d'étiquettes étaient dissimulés dans l'exposition, notamment à la borne de jeu, au passage devant des paparazzi fictifs (prise de photo), et à la borne de sortie. Reliés au système d'information centralisé du musée, ils permettaient, en croisant les données, de faire ressentir au visiteur un réel sentiment d'avoir été « traqué ».

L'objectif pour notre équipe étant le passage à l'échelle, nous avons choisi d'utiliser des technologies moins innovantes, mais beaucoup plus robustes que celles mises en œuvre pour l'expérimentation E-skiing. Nous avons réutilisé l'infrastructure disponible pour l'exposition, en la complétant. Pour cela, nous avons ajouté un lecteur d'étiquettes RFID à l'entrée de l'exposition, de manière à obtenir un horodatage fiable du point d'entrée. Un autre lecteur aurait dû se trouver en sortie de l'exposition et encore d'autres à certains points de passage. Cependant, les contraintes techniques et les aspects liés à la sécurité des personnes (largeur minimum des sorties) ne nous ont pas permis de les installer. Le système d'information du musée a également été modifié de manière à permettre l'enregistrement et le téléchargement journalier des données. Les consignes données pour cette expérimentation étaient réduites au minimum. Les caissières à

l'entrée du Musée devaient simplement donner une carte par visiteur en même temps que le ticket d'entrée. La carte elle-même portait juste l'indication : « gardez cette carte jusqu'à la sortie de l'exposition… elle vous servira. ».

La technologie des étiquettes RFID avait déjà été mise en œuvre dans plusieurs musées afin d'enrichir la visite d'une exposition (par exemple [Hsi & Fait, 2005]). Cependant, à notre connaissance, cette technologie n'avait pas encore été utilisée dans le but d'évaluer de manière systématique l'usage fait par le visiteur de l'environnement ubiquitaire mis à sa disposition. Ainsi, nous avons, tout comme dans l'expérimentation E-skiing, appliqué la technique du cheval de Troie en détournant un dispositif présent dans l'environnement pour s'en servir comme dispositif de capture d'information sur le comportement de l'utilisateur. Des données à caractère personnel étant enregistrées, une déclaration à la CNIL a été effectuée. La CNIL était également partenaire de l'exposition.

Principaux résultats

L'expérimentation s'est déroulée sans trop de difficultés d'ordre technique. Précisons néanmoins que l'évaluation, qui devait durer initialement trois mois, a été réduite aux trois dernières semaines de l'exposition du fait de retards dus aux réglages des lecteurs et à l'impression des cartes contenant les étiquettes RFID. Cela nous a confirmé que les aspects logistiques en situation réelle sont un véritable défi. La limitation technique la plus contraignante a été la faible distance de détection pratique des étiquettes RFID, loin des performances supposées.

L'analyse des données s'est tout d'abord focalisée sur le filtrage des données incohérentes dues soit à des défaillances techniques, soit au non-respect de consignes lors de la distribution des cartes. Ensuite, un premier niveau d'analyse nous a permis de déterminer l'affluence de l'exposition : sur la période des trois semaines de l'expérimentation, 492 cartes ont été distribuées à des visiteurs. Un second niveau d'analyse a eu pour objectif de déterminer les temps de parcours et les types de parcours. Nous avons ainsi pu déterminer que près de la moitié des visiteurs ont effectué le parcours complet tel qu'il était prévu par le scénariste (cf. figure 2). Les résultats détaillés de cette expérimentation peuvent être consultés dans [Jambon, Mandran et al., 2007].

Les biais potentiels liés au protocole expérimental ont été plus délicats à gérer, car nous n'avons pu nous apercevoir de leur existence qu'au moment de dépouiller les données. Dans le protocole expérimental en situation réelle, l'utilisateur ne reçoit pas de consignes spécifiques à l'évaluation, il n'y a donc pas de biais lié à cette consigne. Cependant, les personnels du musée peuvent apporter involontairement des biais. Par exemple, une des personnes de l'accueil n'a pas compris les consignes et n'a pas distribué de carte à tous les visiteurs. Ainsi, les données obtenues certains jours ne recensent pas l'ensemble des comportements des visiteurs. L'échantillon final que nous avons analysé peut donc éventuellement être biaisé, mais nous n'avons pas de moyen de contrôle ni d'estimation de ce biais.

Lors de l'analyse des données, l'absence d'enregistrements vidéo a rendu délicate l'interprétation des anomalies dans les parcours. Dans certains cas, ce sont les commentaires des personnels du musée, notamment les gardiens, qui nous ont permis de découvrir les raisons des comportements insolites détectés. Ce problème est générique à toute expérimentation en situation réelle lorsque l'on ne dispose pas de vidéo. En effet, le faible niveau sémantique des traces recueillies peut rendre difficile l'interprétation

des comportements, car ce travail d'interprétation repose sur la capacité à inférer un comportement à partir des actions de bas niveau. Ceci se produit notamment dans les cas où ce comportement de l'utilisateur n'a pas du tout été anticipé.

Synthèse

Cette expérimentation a montré qu'obtenir des informations statistiquement exploitables en situation réelle est tout à fait envisageable, même sur une large échelle. Elle a également confirmé la complexité technique et logistique de ce type d'expérimentation.

Il était initialement prévu, dans le protocole d'expérimentation, d'analyser l'utilisabilité des bornes, grâce aux traces d'interactions enregistrées en même temps que les parcours, et selon une méthode proche de celle utilisée pour MapMobile. Le musée était principalement intéressé par les parcours, et faute de temps, cette étude n'a été que partiellement réalisée.

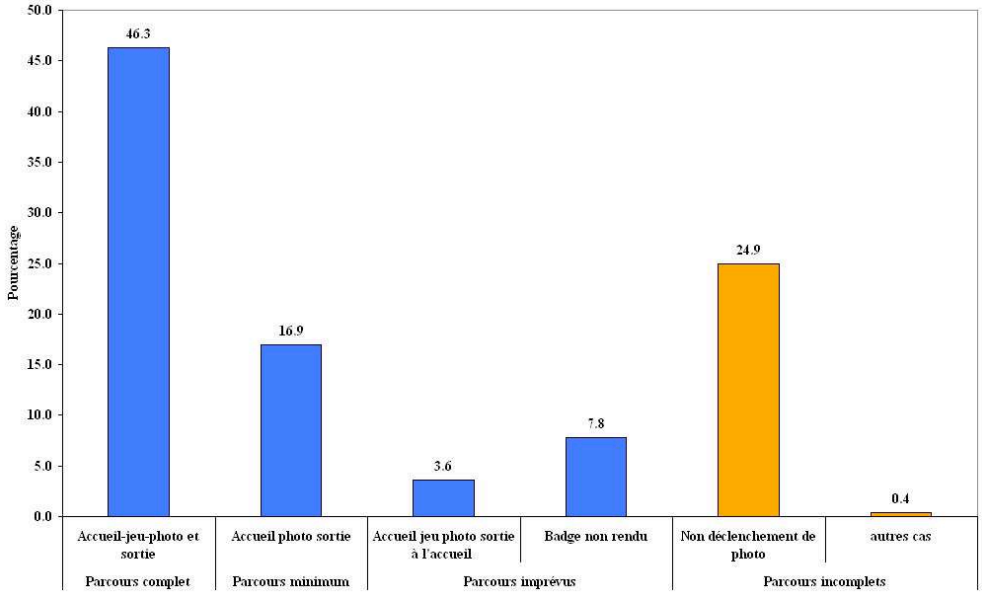


Figure 2 : Types de parcours empruntés par les visiteurs de l'exposition « ni vu – ni connu ».

5. DISCUSSION ET PROPOSITION DE TAXONOMIE

5.1. Retour d'expérience concernant la situation réelle

Les expérimentations en situation réelle nous ont apporté des résultats parfois surprenants, au sens où ils n'étaient pas anticipés, mais de ce fait aussi très pertinents. C'est probablement là leur principal intérêt : mettre en lumière des comportements non attendus des utilisateurs liés à leur activité réelle ou au contexte réel. Ces résultats ont été parfois aussi déroutants et même décevants comparés à l'investissement réalisé. Par exemple, le service E-skiing avait demandé des développements complexes destinés à le rendre compatible avec le réseau de téléphonie mobile, alors qu'au final, une simple borne WiFi associée à un portable situé près du bar aurait suffit… Mais personne, dans

le projet, n'avait anticipé cet usage a priori, bien que cela puisse paraître évident a posteriori.

Nous avons mis en lumière que la détection des problèmes est relativement aisée, mais que leur explicitation requiert toujours la recherche d'informations complémentaires, soit auprès des utilisateurs eux-mêmes en leur faisant revivre une partie de l'expérimentation, soit auprès des témoins naturellement présents dans le contexte lors des expérimentations. A contrario, les consignes sont souvent très aisées à définir car elles sont réduites à leur plus simple expression : une explication du fonctionnement du système.

Nous nous sommes aperçus d'une contrainte du protocole : une fois l'expérimentation lancée, il est très difficile de récupérer les défaillances techniques ou les erreurs commises dans le protocole. Si quelque chose a été oublié, c'est souvent trop tard car on ne peut pas modifier un dispositif ou le protocole sans risquer de perturber le contexte de l'expérimentation et ainsi introduire un biais. Il faut soit laisser faire en acceptant une dégradation des conditions expérimentales, soit prendre la décision d'arrêter toute l'expérimentation. Ce point est important car les expérimentations en situation réelle sont souvent complexes, et donc requièrent d'importants moyens matériels et humains.

En outre, les aspects techniques et logistiques sont complexes, mais tout à fait gérables à condition d'avoir une organisation adaptée et une bonne gestion du temps. Dans tous les cas, les expérimentations en situation réelle requièrent beaucoup plus d'efforts que des expérimentations en laboratoire, avec un différentiel que nous estimons à un facteur deux à dix selon le contexte. Bien entendu, de nombreux tests, dont au moins un en situation quasiment identique, sont nécessaires. Il est illusoire de croire qu'il est possible de tout prévoir sur le papier, car l'intérêt de la situation réelle est justement l'occurrence des imprévus.

Enfin, il est primordial de toujours enregistrer le plus de données possible, même a priori redondantes ou secondaires. En effet, il se produira probablement des pannes des systèmes d'enregistrement ou plus probablement il apparaitra des situations inédites, et il faudra reconstituer ces situations à partir d'indices issus des données que l'on avait jugées, au départ, secondaires.

5.2. Généralisation : les quatre approches

En prenant comme point de départ la distinction entre laboratoire et terrain, usuelle dans la littérature, nous avons proposé la notion de situation réelle comme une spécialisation du terrain. Or, la recherche d'une classification de ces configurations nous a amenés à définir une autre configuration, la « simulation réaliste ». En outre, nous avons associé à l'ensemble de ces configurations, de nouvelles appellations, basées sur une terminologie inspirée de la biologie, plus précise et, de plus, cohérente avec les propositions de Kjeldskov et al. [Kjeldskov & Skov, 2007].

Le laboratoire (in-vitro)

Le laboratoire d'utilisabilité est la configuration de référence. Il se caractérise notamment par un environnement artificiel contrôlé. Ainsi, les problèmes d'utilisabilité détectés seront directement dépendants des tâches effectuées par l'utilisateur, et donc des scénarios choisis. Il est donc peu probable que des éléments liés à l'usage réel seront découverts.

Nous nommons désormais cette configuration « in-vitro » en accord avec les propositions de Kjeldskov et al. [Kjeldskov & Skov, 2007].

Le terrain (in-situ)

Comparé au laboratoire d'utilisabilité, l'intérêt principal du terrain est d'apporter les éléments du contexte qui vont rendre l'évaluation plus pertinente. Le protocole expérimental étant peu différent du laboratoire d'utilisabilité, notamment la tâche de l'utilisateur n'étant le plus souvent pas libre, les problèmes ergonomiques détectés dépendront directement du choix des scénarios donnés dans la consigne. Ainsi, il n'est pas surprenant que les problèmes détectés sur le terrain soient similaires en type et en nombre dans ces deux situations. De plus, la présence éventuelle du facilitateur apporte un biais connu dans le déroulement de l'activité du sujet.

Si l'environnement est peu contraint, le terrain peut apporter son lot d'événements permettant de rendre l'évaluation un peu plus réaliste, même si le sujet est souvent isolé du monde extérieur par la « bulle » formée par la présence des observateurs. C'est donc avant tout des problèmes d'utilisabilité que l'on va détecter, mais resitués dans le contexte d'usage réel du système. Par exemple, si un écran de téléphone mobile est inadapté à la luminosité extérieure, ce problème ne sera pas détecté en laboratoire d'utilisabilité, mais le sera sur le terrain, à condition qu'il fasse beau… C'est là l'une des limitations de cette configuration. En effet, les éléments du contexte étant par nature non prévisibles, rien ne garantit qu'ils se produiront. Le risque est alors d'obtenir des résultats biaisés car certaines variables indépendantes ne sont pas prises en compte.

Nous nommons désormais cette configuration « in-situ » en accord avec les propositions de Kjeldskov et al. [Kjeldskov & Skov, 2007].

La situation réelle (in-vivo)

Concernant les expérimentations en situation réelle, le protocole expérimental change radicalement. En effet, le facilitateur ne peut plus intervenir et l'utilisateur n'a, a priori, que des tâches libres. Il va donc « vivre sa vie » avec le système. Deux types de résultats seront obtenus : des statistiques d'usage du dispositif, c'est-à-dire les tâches que l'utilisateur a souhaité effectuer, et l'utilisabilité « filtrée par l'usage », c'est-à-dire les problèmes ergonomiques liés aux tâches effectuées dans le contexte d'usage. Ces tâches étant un sous-ensemble des tâches possibles, il est cohérent que le nombre de problèmes ergonomiques détectés en situation réelle soit bien moindre qu'en laboratoire d'utilisabilité ou que sur le terrain. Néanmoins, les problèmes ergonomiques détectés gagnent en pertinence car ils ont été mis en lumière par l'usage réel du dispositif.

Cependant, les résultats doivent être pris avec circonspection, car les tâches effectuées par les sujets et les événements du contexte ne se produisent pas de manière déterministe, et constituent ainsi des variables indépendantes masquées. Pour minimiser ce risque, il peut être intéressant, si cela est réalisable, d'augmenter significativement le nombre de sujets et le temps d'expérimentation, en espérant que les événements statistiquement probables se produiront. De plus, le fait même que l'utilisateur ait connaissance de faire partie d'une évaluation apporte a priori un biais, connu sous le nom de « effet Hawthorne » [Macefield, 2007]. Cet effet, bien que très controversé, suggère que la performance des utilisateurs s'améliore du fait même qu'ils aient conscience de faire partie d'une expérimentation.

Nous nommons désormais cette configuration « in-vivo» afin d'être cohérent avec les autres appellations. Remarquons que Kjeldskov et al. ne mentionnent pas cette configuration dans leur article [Kjeldskov & Skov, 2007].

La quatrième configuration : la simulation réaliste (in-simu)

Dans l'inspiration du « Truman Show »[8], il existe une quatrième configuration. Celle-ci consiste à avoir un contexte artificiel, mais réaliste, et des tâches pouvant être libres, prescrites, ou encore basées sur des scénarios réalistes. Cette configuration ne peut pas être adaptée à toutes les situations car il est nécessaire de pouvoir simuler correctement le contexte, et la durée des tâches doit être compatible avec l'usage du simulateur. Deux auteurs ont exploré cette configuration sous des appellations très différentes. Kjeldskov et al. sous l'appellation « in-sitro » [Kjeldskov & Skov, 2007] et Hertzum sous l'appellation « workshop » [Hertzum, 1999]. Dans ces deux publications, les résultats obtenus en simulation se rapprochent de ceux obtenus sur le terrain.

Cette approche est à rapprocher des nouvelles techniques de type LOFT (« Line Oriented Flight Training ») utilisées en formation et entrainement des pilotes de l'aviation civile, consistant à simuler non pas uniquement des séries de situations d'urgence, mais des parties de vol réalistes où se mélangent situations normales, situations exceptionnelles et situations d'urgence. D'un certain point de vue, cela peut également se rapprocher de la notion de « living laboratory » dont la définition exacte n'est pas encore bien stabilisée. Pour nous, nous considérons cette configuration comme une situation réelle dégradée où le contexte est en partie contraint.

Nous nommons désormais cette configuration « in-simu» afin d'être cohérent avec les autres appellations. Kjeldskov et al. mentionnent cette configuration sous le nom de « in-sitro » [Kjeldskov & Skov, 2007], car elle y est présentée comme intermédiaire entre « in-vitro » et « in-situ ». Nous n'avons pas conservé ce terme car nous l'avons jugé peu explicite, et de plus, notre classification est différente et ne situe pas cette configuration à la même place.

5.3. Proposition de classification

Considérant les quatre configurations possibles, la tâche (prescrite, scénario réaliste ou libre) et le contexte (artificiel, simulé ou réel) semblent se dégager comme les deux principales grandeurs orthogonales. Nous avons masqué les aspects liés au réalisme du dispositif et à la présence des observateurs et dispositifs d'observation car, liés aux deux autres, ils nous ont semblés secondaires. En effet, le réalisme du dispositif n'est véritablement intéressant que lorsqu'il va de pair avec le réalisme du contexte, et la présence des observateurs n'est pas souhaitable lorsque le contexte est réel ou les tâches libres.

Nous déduisons ainsi neuf situations théoriquement possibles (figure 3). Nous les avons regroupées selon les quatre configurations définies précédemment: in-vitro, in-simu, in-situ et in-vivo. Globalement, c'est le réalisme du contexte qui détermine la configuration, avec pour le contexte réel, une distinction selon que les tâches sont libres ou non.

[8] « The Truman Show » est un film relatant une émission de télé-réalité où le sujet vit, à son insu, dans un monde totalement artificiel où toutes les personnes qu'il rencontre sont des acteurs et où tous les événements se produisant sont déclenchés par un réalisateur.

A l'intérieur de chaque configuration, toutes les situations sont théoriquement possibles. Cependant, nous avons distingué les situations qui nous paraissaient les plus intéressantes (cercles noirs), celles intermédiaires (cercles gris) et celles qui nous semblaient moins intéressantes (cercles blancs).

Les situations les moins intéressantes sont les situations extrêmes où le protocole expérimental nous a semblé peu réalisable. Ainsi, l'usage de tâches libres en contexte artificiel nous a semblé peu pertinent car il peut être difficile pour un utilisateur d'agir de manière « naturelle » dans le contexte très contraint d'un laboratoire d'utilisabilité. De même, demander à un utilisateur d'exécuter scrupuleusement un ensemble de tâches prescrites en contexte réel, compte tenu des sollicitations externes, nous a semblé peu réaliste.

A l'opposé, nous pensons que les situations les plus intéressantes sont les situations qui présentent une bonne adéquation entre la prescription des tâches et le réalisme du contexte. En effet, le contexte très artificiel d'un laboratoire s'adapte bien à des tâches prescrites, le contexte simulé permet aux utilisateurs de suivre des scénarios réalistes dans un contexte proche de la réalité, tandis que les tâches libres ne se justifient vraiment qu'en contexte réel.

Les situations intermédiaires sont un compromis entre les deux précédentes. Remarquons que la configuration in-situ ne comporte pas de situation parmi les plus intéressantes. En effet, nous pensons que cette configuration n'est pas intéressante en elle-même, mais peut être un bon compromis lorsque les configurations in-simu ou in-vivo ne peuvent être mises en œuvre.

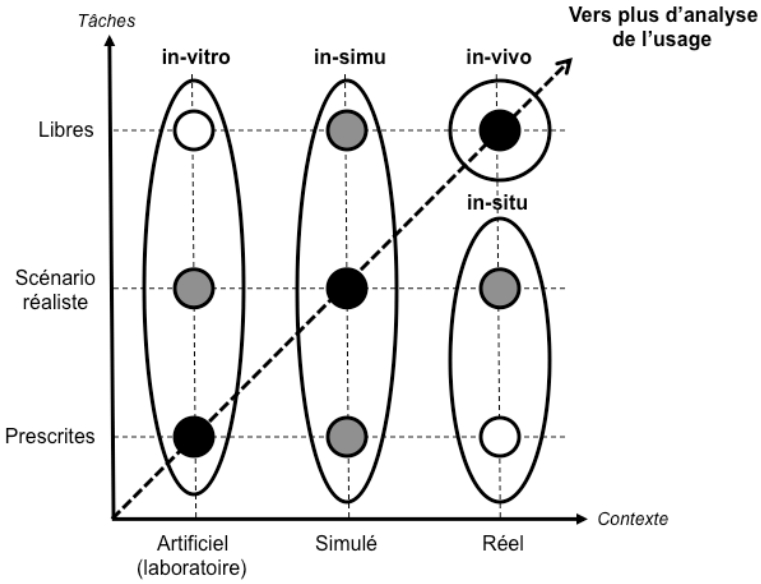


Figure 3 : Typologie des configurations en fonction du contexte et des tâches.

5.4. Quelle démarche adopter ?

Les quatre configurations définies ci-dessus nous semblent plus complémentaires qu'opposées. De notre point de vue, la configuration in-vitro permet d'étudier en détail les problèmes ergonomiques posés par le système interactif, selon des scénarios choisis.

Cette configuration se focalise sur l'utilisabilité. La configuration in-simu permet de détecter les fonctionnalités qui font sens pour l'utilisateur ainsi que certains problèmes d'utilisabilité, mais dans un environnement contrôlé. Enfin, la configuration in-vivo permet de finaliser l'étude en insistant sur les fonctionnalités qui font sens pour l'utilisateur dans le contexte réel où il est. Cette dernière configuration se focalise sur l'usage. Cas particulier, la configuration in-situ peut se substituer aux configurations in-simu ou in-vivo si, respectivement, le contexte est difficile à simuler ou si laisser l'utilisateur effectuer des tâches libres n'a pas de sens. Le choix d'une configuration pourra aussi être déterminé par les objectifs de l'évaluation ou par les limites techniques du système testé. En effet, in-vitro, il est possible d'utiliser des maquettes (par exemple via la technique du Magicien d'Oz), ce qui est quasiment impossible in-vivo où le système doit être perçu comme pleinement opérationnel (à de rares exception près [Consolvo et al., 2007]).

L'ordre « naturel » que nous proposons pour un dispositif donné consiste à effectuer les trois types d'évaluation dans l'ordre {in-vitro → in-simu → in-vivo} de manière à disposer progressivement d'un système dont les problèmes ergonomiques auront été éliminés avant de s'intéresser à son usage. Cette approche présente néanmoins un défaut : si certaines des fonctionnalités ne font pas sens à l'utilisateur, elles auront néanmoins été mises au point, à grands frais, lors des deux étapes précédentes. Il est possible aussi de prendre le contre-pied de cet ordre « naturel » et de suivre un ordre inverse : {in-vivo → in-vitro}. L'idée ici est de détecter les usages qui font sens en amont, avant de se concentrer ensuite sur leur mise au point. Le défaut de cette approche est qu'un problème ergonomique bloquant pour l'utilisateur va donner de faux résultats négatifs sur l'usage, c'est-à-dire qu'une fonctionnalité inutilisable ne sera tout simplement pas utilisée... De manière générale, il est également possible d'utiliser les quatre configurations de manière opportuniste, selon un cycle de développement plus général comme nous l'avons proposé sous la forme des « Moments de la conception » [Caelen, Jambon, & Vidal, 2005]. Ici, chaque type d'évaluation est un « Moment ». Libre aux concepteurs, en fonction de leurs besoins et de l'évolution du projet, de choisir les configurations et l'ordre dans lequel elles seront utilisées.

6. CONCLUSION ET PERSPECTIVES

Ces travaux nous ont permis de mettre en lumière la pertinence des expérimentations in-vivo, notamment lorsqu'elles sont associées à la technique du « cheval de Troie » qui permet la capture de données comportementales en minimisant les biais liés à l'observation. Cette configuration nous semble plus adaptée à détecter les problèmes liés à l'usage du dispositif qu'à son utilisabilité, où d'autres techniques, comme les expérimentations in-vitro, in-simu ou in-situ, moins complexes à mettre en œuvre, peuvent être utilisées. Nous avons également proposé une taxonomie des configurations d'expérimentation qui peut être vue selon deux dimensions principales : la tâche et le contexte. Ces travaux ont été effectués dans le cadre des études sur les systèmes mobiles et ubiquitaires, car ces systèmes requièrent une immersion dans leur contexte d'usage, mais ils sont probablement applicables dans le cadre des systèmes sédentaires plus classiques.

Actuellement, les perspectives de notre travail s'orientent vers deux directions. Premièrement, nous cherchons à affiner notre connaissance des différences entre les types de configurations, en s'intéressant à élargir la liste de nos critères de comparaison. Des pistes intéressantes peuvent être trouvées dans les travaux comparant les conditions

expérimentales, issus du domaine de l'analyse de comportements en économie (par exemple les travaux de Fiedler et al. [Fiedler & Haruvy, 2009]) ou encore du domaine des systèmes d'apprentissage basés sur la simulation (par exemple les travaux Dahl et al. [Dahl, Alsos, & Svanæligs, 2010]). Nous nous intéressons également à formaliser et rendre plus reproductible les comparaisons entre les configurations expérimentales, par exemple en se basant sur les critères de comparaison proposés par Hartson et al. [Hartson, Andre, & Williges, 2003], inspirés du domaine de la recherche d'information. En corolaire, les expérimentations avec tâches libres sont très souvent synonymes de très grandes quantités de données à traiter, et, qui plus est, des données de faible niveau d'abstraction qu'il faut analyser. Notre seconde perspective est d'automatiser l'analyse de ces données. Les projets TELEOS et TRACES qui s'intéressent respectivement à l'analyse des gestes chirurgicaux et à l'analyse des déplacements des personnes, en sont actuellement les supports.

REMERCIEMENTS

Ces expérimentations ont été financées conjointement par les projets RNTL franco-finlandais ADAMOS et Région Rhône-Alpes IMERA. Outre les auteurs de l'article, de nombreuses personnes se sont investies dans la mise en œuvre des expérimentations, à la fois du point de vue technique et du point de vue logistique. Les auteurs tiennent ainsi à remercier toutes les personnes ayant contribué au succès de ces trois expérimentations.

Pour MapMobile : D. David du CEA-Leti, A. Martin de France Télécom R&D, F. Forest de la MSH-Alpes, C. Golanski, M. Pernière et P.-J. Pommier du LIG/MultiCom (ex. CLIPS-IMAG). Pour E-skiing : P. Schermesser, A. Vidal et D. David du CEA-Leti, F. Forest de la MSH-Alpes, M. Léger et D. Maréchal de Rossignol ainsi que la station de l'Alpe d'Huez pour son accueil. Pour l'exposition « ni vu – ni connu » au Muséum du département du Rhône : C. Sermet, le scénographe de l'exposition, J.-F. Salmon le responsable informatique du Muséum, ainsi que le centre ERASME et la société Tagproduct qui en ont réalisé la mise en œuvre technique.

REFERENCES

Arhippainen, L., Rantakokko, T., & Tähti, M. 2004. *Mobile Feedback Application for Emotion and User Experience Collection.* Paper presented at the Proactive computing workshop - PROW 2004, Helsinki, Finland.

Baillie, L., & Schatz, R. 2005. *Exploring multimodality in the laboratory and the field.* Paper presented at the Proceedings of the 7th International Conference on Multimodal Interfaces, Torento, Italy.

Bartek, V., & Cheatham, D. 2003. *Experience remote usability testing, Part 1 & 2* (URL: "http://www-106.ibm.com/developerworks/web/library/wa-rmusts1/"): IBM's Pervasive Computing Division.

Bastien, C., & Scapin, D. 1995. Evaluating a user interface with ergonomic criteria. *International Journal of Human-Computer Interaction, 7*(2), 105-121.

Betiol, A. H., & Cybis, W. d. A. 2005. *Usability Testing of Mobile Devices: A Comparison of Three Approaches.* Paper presented at the IFIP TC13 International Conference on Human-Computer Interaction - Interact'2005, Rome, Italy.

Brewster, S., & Walker, A. 2000. *Non-Visual Interfaces for Wearable Computers.* Paper presented at the IEE workshop on Wearable Computing (IEE, 00/145), London.

Brush, A. J. B., Ames, M., & Davis, J. 2004. *A comparison of synchronous remote and local usability studies for an expert interface.* Paper presented at the CHI'04

conference on Human factors in computing systems extended abstracts, Vienna, Austria.

Buisson, M., Bustico, A., Chatty, S., Colin, F.-R., Jestin, Y., Maury, S., et al. 2002. *Ivy: un bus logiciel au service du développement de prototypes de systèmes interactifs.* Paper presented at the Proceedings of the 14th French-speaking conference on Human-computer interaction (Conférence Francophone sur l'Interaction Homme-Machine), Poitiers, France.

Caelen, J., Jambon, F., & Vidal, A. 2005. Conception participative : des "Moments" à leur instrumentation. *Revue d'Interaction Homme-Machine (RIHM), 6*(2), 1-29.

Calvet, G., Salembier, P., Kahn, J., & Zouinar, M. 2005. *Étude empirique de l'interaction multimodale en mobilité : approche méthodologique et premiers résultats.* Paper presented at the 17e Conférence Francophone sur l'Interaction Homme-Machine - IHM'05, Toulouse, France.

Castillo, J. C., Hartson, H. R., & Hix, D. 1998. *Remote usability evaluation: can users report their own critical incidents?* Paper presented at the CHI'98 conference on Human factors in computing systems, Los Angeles, California, United States.

Chapanis, A. 1967. The Relevance of Laboratory Studies to Practical Situations. *Ergonomics, 10*(5), 557 - 577.

Consolvo, S., Harrison, B., Smith, I., Chen, M. Y., Everitt, K., Froehlich, J., et al. 2007. Conducting In Situ Evaluations for and With Ubiquitous Computing Technologies. *International Journal of Human-Computer Interaction, 22*(1), 103-118.

Dahl, Y., Alsos, O. A., & Svanæligs, D. 2010. Fidelity Considerations for Simulation-Based Usability Assessments of Mobile ICT for Hospitals. *International Journal of Human-Computer Interaction, 26*(5), 445-476.

Demumieux, R., & Losquin, P. 2005. *Collecter les usages réels des clients de téléphonie mobile (un outil embarqué).* Paper presented at the 17e Conférence Francophone sur l'Interaction Homme-Machine - IHM'05, Toulouse, France.

Duh, H. B.-L., Tan, G. C. B., & Chen, V. H.-H. 2006. *Usability evaluation for mobile device: a comparison of laboratory and field tests.* Paper presented at the Proceedings of the 8th conference on human-computer interaction with mobile devices and services, Helsinki, Finland.

Fiedler, M., & Haruvy, E. 2009. The lab versus the virtual lab and virtual field – An experimental investigation of trust games with communication. *Journal of Economic Behavior & Organization, 72*(2), 716-724.

Fields, B., Amaldi, P., Wong, W., & Gill, S. 2007. In Use, In Situ: Extending Field Research Methods *International Journal of Human-Computer Interaction, 22*(1&2), 1-6.

Forest, F., Guilloux, V., Mallein, P., & Panisset, J. 1998. La conception assistée par l'usage dans les bibliothèques publiques. In M.-H. Kœnig (Ed.), *Connaître ses publics : savoir pour agir (La Boîte à outils)* (Vol. 8). Villeurbanne: Institut de Formation des Bibliothécaires - IFB.

Froehlich, J., Chen, M. Y., Consolvo, S., Harrison, B., & Landay, J. A. 2007. *MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones.* Paper presented at the Proceedings of the 5th International Conference on Mobile Systems, Applications and Services (MobySys'07), San Juan, Puerto Rico.

Gob, A., & Drouguet, N. 2006. *La muséologie. Histoire, développement, enjeux actuels*: Armand Colin.

Goodman, J., Brewster, S., & Gray, P. 2004. *Using Field Experiments to Evaluate Mobile Guides.* Paper presented at the 3rd Annual Workshop on HCI in Mobile Guides, Glasgow, UK.

Hagen, P., Robertson, T., Kan, M., & Sadler, K. 2005. *Emerging research methods for understanding mobile technology use.* Paper presented at the OzCHI'2005, Canberra, Australia.

Hammontree, M., Weiler, P., & Nayak, N. 1994. Remote Usability Testing. *interactions, 1*(3), 21-25.

Hartson, H. R., Andre, T. S., & Williges, R. C. 2003. Criteria For Evaluating Usability Evaluation Methods. *International Journal of Human-Computer Interaction, 15*(1), 145-181.

Hartson, H. R., Castillo, J. C., Kelso, J., & Neale, W. C. 1996. *Remote evaluation: the network as an extension of the usability laboratory.* Paper presented at the Conference on human factors in computing systems (CHI'96), Vancouver, British Columbia, Canada.

Hertzum, M. 1999. *User Testing in Industry: A Case Study of Laboratory, Workshop, and Field Tests.* Paper presented at the 5th ERCIM Workshop on "User Interfaces for All" – Special Theme "User-Tailored Information Environments", Dagstuhl, Germany.

Hsi, S., & Fait, H. 2005. RFID enhances visitors' museum experience at the Exploratorium. *Communicatons of the ACM, 48*(9), 60-65.

Isomursu, M., Kuutti, K., & Väinämö, S. 2004. *Experience clip: method for user participation and evaluation of mobile concepts.* Paper presented at the 8th conference on Participatory design: Artful integration: interweaving media, materials and practices, Toronto, Ontario, Canada.

Jambon, F. 2006. *Retours d'expérience lors d'évaluation de dispositifs mobiles en situation réelle.* Paper presented at the 3e Journées Francophones Mobilité et Ubiquité - Ubimob'06, Paris, France.

Jambon, F. 2009. User Evaluation of Mobile Devices: In-Situ versus Laboratory Experiments. *International Journal of Mobile Computer-Human Interaction (IJMHCI), 1*(2), 56-71.

Jambon, F., Golanski, C., & Pommier, P.-J. 2006. *Évaluation des dispositifs mobiles : sur le terrain ou en laboratoire ?* Paper presented at the 18e Conférence Francophone sur l'Interaction Homme-Machine - IHM'06, Montréal, Canada.

Jambon, F., Golanski, C., & Pommier, P.-J. 2007. *Meta-Evaluation of a Context-Aware Mobile Device Usability.* Paper presented at the International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM'07).

Jambon, F., Mandran, N., Meillon, B., & Perrot, C. 2008. *Évaluation des systèmes mobiles et ubiquitaires : proposition de méthodologie et retours d'expérience.* Paper presented at the Ergo-IA'08 "L'humain au coeur des systèmes et de leur développement", Biarritz, France.

Jambon, F., Mandran, N., & Perrot, C. (2007). La RFID au service de l'analyse du parcours muséal des visiteurs. *La lettre de l'Office de Coopération et d'Information Muséographiques (OCIM),* 11-17.

Jambon, F., & Meillon, B. 2009. *User Experience Evaluation in the Wild.* Paper presented at the CHI'09 Extended Abstracts on Human Factors in Computing Systems, Boston, MA, USA.

Jensen, K. L., & Larsen, L. B. 2007. *Evaluating the usefulness of mobile services based on captured usage data from longitudinal field trials.* Paper presented at the 4th

international conference on mobile technology, applications, and systems and the 1st international symposium on Computer human interaction in mobile technology, Singapore.

Johnson, H. H., Solso, R. L., & Beale, K. 1997. Ethics of Experimental Research. In *Experimental Psychology: A Case Approach (6th Revised edition)* (pp. 400): Addison Wesley Longman.

Kaikkonen, A., Kekäläinen, A., Cankar, M., Kallio, T., & Kankainen, A. 2005. Usability Testing of Mobile Applications: A Comparison between Laboratory and Field Testing. *Journal of Usability Studies, 1*(1), 4-17.

Kellar, M., Reilly, D., Hawkey, K., Rodgers, M., MacKay, B., Dearman, D., et al. 2005. *It's a jungle out there: practical considerations for evaluation in the city.* Paper presented at the CHI'05 extended abstracts on Human factors in computing systems, Portland, OR, USA.

Kjeldskov, J., & Graham, C. 2003. *A Review of Mobile HCI Research Methods.* Paper presented at the Mobile Human-Computer Interaction – MobileHCI'2003, Udine, Italy.

Kjeldskov, J., Graham, C., Pedell, S., Vetere, F., Howard, S., Balbo, S., et al. 2005. Evaluating the Usability of a Mobile Guide: The Influence of Location, Participants and Resources. *Behaviour & Information Technology, 24*(1), 51-65.

Kjeldskov, J., Skov, M., Als, B., & Høegh, R. 2004. *Is It Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field.* Paper presented at the Mobile Human-Computer Interaction - MobileHCI'2004, Glasgow, UK.

Kjeldskov, J., & Skov, M. B. 2007. Studying Usability In Sitro: Simulating Real World Phenomena in Controlled Environments. *International Journal of Human-Computer Interaction, 22*(1), 7-36.

Kjeldskov, J., & Stage, J. 2004. New Techniques for Usability Evaluation of Mobile Systems. *International Journal of Human-Computer Studies - IJHCS, 60*(5-6), 599-620.

Lyons, K., & Starner, T. 2001. *Mobile Capture for Wearable Computer Usability Testing.* Paper presented at the Fifth International Symposium on Wearable Computers (ISWC'01), Zurich, Switzerland.

Macefield, R. 2007. Usability studies and the Hawthorne Effect. *Journal of Usability Studies, 2*(3), 145-154.

Malaurie, J. (Ed.). 2002. *De la vérité en ethnologie...* (Polaire ed.): Economica.

Molich, R. 2000. *Usable Web Design (In Danish)*: Ingeniøren|bøger.

Pirhonen, A., Brewster, S., & Holguin, C. 2002. *Gestural and audio metaphors as a means of control for mobile devices.* Paper presented at the ACM-SIGCHI Conference on Human Factors in Computing Systems (CHI'2002), Minneapolis, Minnesota, USA.

Po, S., Howard, S., Vetere, F., & Skov, M. B. 2004. *Heuristic Evaluation and Mobile Usability: Bridging the Realism Gap.* Paper presented at the Mobile Human-Computer Interaction (MobileHCI'2004), Glasgow, UK.

Riegelsberger, J., & Nakhimovsky, Y. 2008. Seeing the bigger picture: a multi-method field trial of google maps for mobile, *CHI'08 extended abstracts on Human factors in computing systems* (pp. 2221-2228). Florence, Italy: ACM Press.

Roto, V., Oulasvirta, A., Haikarainen, T., Lehmuskallio, H., & Nyyssönen, T. 2004. *Examining mobile phone use in the wild with quasi-experimentation* (Technical

Report No. 2004-1). Helsinky, Finland: Helsinky Institute for Information Technology (HIIT).

Rowley, D. E. 1994. *Usability testing in the field: bringing the laboratory to the user.* Paper presented at the Conference on Human Factors in Computing Systems (CHI'94), Boston (MA), United States.

Scholtz, J. 2001. *Adaptation of Traditional Usability Testing Methods for Remote Testing.* Paper presented at the 34th Annual Hawaii International Conference on System Sciences - HICSS'01, Maui, Hawaii.

Schulte-Mecklenbeck, M., & Huber, O. 2003. Information search in the laboratory and on the Web: With or without an experimenter. *Behavior Research Methods, Instruments & Computers, 35*(2), 227-235.

Sermet, C., & Millet, M. 2007. "Ni vu, ni connu", une scénographie de camouflages. *La lettre de l'OCIM (Office de Coopération et d'Information Muséographiques)*(113), 4-10.

Thomas, J. C., & Kellogg, W. A. 1989. Minimizing Ecological Gaps in Interface Design. *IEEE Software, 6*(1), 78-86.

Thompson, K. E., Rozanski, E. P., & Haake, A. R. 2004. *Here, there, anywhere: remote usability testing that works.* Paper presented at the 5th conference on Information Technology Education, Salt Lake City, UT, USA.

Tullis, T., Fleischman, S., McNulty, M., Cianchette, C., & Bergel, M. 2002. *An Empirical Comparison of Lab and Remote Usability Testing of Web Sites.* Paper presented at the Usability Professionals Association Conference - UPA'2002, Orlando, FL, USA.

Waterson, S., Landay, J. A., & Matthews, T. 2002. *In the lab and out in the wild: remote web usability testing for mobile devices.* Paper presented at the CHI'02 conference on Human Factors in Computing Systems extended abstracts, Minneapolis, Minnesota, USA.

Zouinar, M., Relieu, M., Salembier, P. & Calvet, G. 2004. *Observation et capture de données sur l'interaction multimodale en mobilité.* Paper presented at the Premières journées francophones Mobilité et Ubiquité - UbiMob'04, Sophia-Antipolis.

Francis JAMBON est Maître de Conférences en informatique à Polytech'Grenoble (Université Joseph Fourier) où il enseigne la modélisation des systèmes d'information, l'interaction homme-machine et la conception des systèmes mobiles. Il fait partie de l'équipe MultiCom du LIG (Laboratoire d'Informatique de Grenoble) depuis 2002. Auparavant, et depuis 1998, il a fait partie de l'équipe IHM et PSE du laboratoire LISI/ENSMA à Poitiers. Ses travaux de recherche portent principalement sur les méthodes d'évaluation et de validation des systèmes interactifs, et d'analyse de l'activité humaine, notamment dans le cas des systèmes mobiles, ubiquitaires ou critiques. Il est membre de l'ACM-SIGCHI et de l'AFIHM.

Nadine MANDRAN est ingénieur d'étude CNRS au Laboratoire d'Informatique de Grenoble (LIG). Après une formation en statistique et en sociologie, elle a exercé ses activités dans plusieurs domaines de recherche. Aujourd'hui, son activité se déroule dans le cadre de la plateforme expérimentale du LIG Marvelig. Elle travaille à la mise en œuvre de méthodologies et de protocoles expérimentaux pour les équipes de recherche dont le domaine est lié aux comportements des utilisateurs face à des nouveaux

concepts. Cet axe permet d'assurer conseil et partenariat méthodologique pour la production et le traitement qualitatif ou statistique des données. Un autre axe est le déploiement de la démarche qualité pour la plateforme expérimentale. Dans ce cadre elle met en place différents processus pour la capitalisation des prototypes du laboratoire et pour le suivi des expérimentations. Elle est membre du Réseau Qualité du CNRS.

Brigitte MEILLON est ingénieur CNRS au LIG (Laboratoire d'informatique de Grenoble), dans l'équipe MultiCom. Elle est responsable technique de la plateforme d'expérimentation située au CTL (Centre des technologies du Logiciel), sur le campus de Grenoble. Son activité est centrée sur l'instrumentation des expérimentations et sur la capture de comportements d'utilisateurs. En fonction des objectifs des différents projets de recherche et contrats, elle propose des outils existants ou développés en interne pour produire des données pertinentes et pour les traiter. Elle fait évoluer le matériel et les outils de la plateforme, en fonction des différents projets et contrats de l'équipe (domotique, géolocalisation, RFID). Elle s'intéresse plus particulièrement à l'oculométrie d'un point de vue technique depuis une dizaine d'années et a développé des logiciels d'acquisition et de traitement de données oculométriques.

Christian PERROT est ingénieur de recherche au CNRS, chargé de valorisation dans l'équipe MultiCom du LIG. Il conduit des projets à caractère industriel autour des RFID. Il est responsable technique du projet de table interactive mettant en jeu des objets Tangibles dans le cadre du projet ANR « TTT ». Il contribue également au déploiement de solutions technologiques robustes dans les musées dans le domaine des outils de médiation fixes ou mobiles associés à des technologies de traçage du comportement des visiteurs.