# A Framework for Remote User Evaluation of Accessibility and Usability of Websites

Christopher Power, Helen Petrie, and Richard Mitchell

Department of Computer Science
University of York
York, YO10 5DD, UK
{cpower,petrie,mitchell}@cs.york.ac.uk

**Abstract.** The inclusion of participants that are representative of the diverse populations of users is essential for meaningful and useful evaluations of usability and accessibility on the web. This paper proposes the requirements and architecture for an automated tool suite to help manage the design and deployment of evaluations to these participants. A prototype implementation of this architecture that is being prepared is also discussed.

## 1 Introduction

There is a need to evaluate designs and prototypes much earlier than is currently common practice. Indeed, there has been an ongoing call in the usability and accessibility communities to increase user testing for web applications; a call that, after a decade of experience with web systems, has not been realized. The current lack of user engagement often stems from the inability to collect large enough samples of representative users to get meaningful results from the evaluations. This problem is magnified when evaluating web applications with people with disabilities due to the variety of user agents and assistive technologies that must be accounted for during the evaluation process. In this case, the combination of technologies and user preferences further divides participants into smaller and smaller subgroups, from each of which data must be collected.

One solution to the challenge of collecting enough evaluation results, particularly about web applications, is to engage users in remote evaluations. These evaluations provide valid data [2, 8] without the logistical problems associated with having users visit a laboratory [3]. Further, it allows users to conduct ecologically valid evaluations in their own homes and places of business. This gives designers a view of how their applications will be used in real environments after deployment and most importantly, an understanding of the impacts that different technologies will have on the web application after deployment.

Currently, many of these remote evaluations are conducted through bespoke implementations designed for testing a single web application, and as a result a great deal time and effort are committed to creating a test environment specifically tailored to that single application. This type of extraneous development contributes to the low adoption of remote evaluations in practice due to the time and resources that designers and developers are reluctant to commit.

C. Stephanidis (Ed.): Universal Access in HCI, Part I, HCII 2009, LNCS 5614, pp. 594–601, 2009.

In order to reduce the resources required for remote user evaluations the evaluators require a robust, usable environment for constructing the evaluation trials for their web application. This paper discusses the design and architecture of an online application intended to meet the requirements of these web practitioners. The paper begins by discussing some of the challenges associated with conducting user evaluations and what is required of a suite of tools or applications to address these challenges. Following this, the authors discuss the architecture and implementation of a prototype of such a suite of tools, named Klingsor, to assist evaluators in managing evaluations and delivering them to targeted user groups. The paper will conclude with a discussion of future work, included pilot tests of the evaluation application scheduled for the autumn of 2009.

## 2 Related Research

Remote evaluation has been used for over a decade, with researchers and practitioners moving the evaluation of interactive technology beyond the walls of the laboratory setting. Barriers to user involvement such as distance of travel for participants, evaluator time and overall cost of evaluation become lessened through the application of remote evaluation methods. Indeed, remote evaluation, particularly task-based asynchronous remote evaluation, also provides for the engagement of a broader range of users in evaluation activities due to the ability of users to administer the trials themselves without being dependent on the presence of an evaluator. As such, web applications can be evaluated with a wide variety of people, technology configurations and environments [11]. This is particularly attractive for purposes of testing interactive systems with people with disabilities, where the wide variety of personal preferences, assistive technology and user agents lead designers, developers and researchers to claim that it is impossible to get a representative sample of users [8].

Due to the potential benefits that can be gained from remote evaluations, a great deal of work has been undertaken in the last decade examining how to exploit the networking and mobility aspects of personal computing to conduct remote evaluations. Hartson, Castillo *et al*. [4,5] identified several different methods for conducting remote evaluations. These remote evaluations were classified in the following subcategories:

- *Portable evaluations* where an evaluation unit conducts evaluations in the users' own work/home environments.
- *Local evaluation at a remote site* where external equipment is installed at a remote site and evaluations are conducted at that remote site.
- *Remote questionnaires/surveys* where the display of the survey to the participant is triggered by actions in the interface.
- *Remote control evaluation* where the users' environments are equipped with recording equipment for synchronous or asynchronous recording of data.
- *Video conferencing as an extension of the usability laboratory*, a technique that involves the users undertaking particular tasks while engaging in synchronous communication with an evaluator at a remote site.

- *Instrumented remote evaluation* where user applications are augmented with components that record information about user workflows.
- *Semi-instrumented remote evaluation* where the users are trained to identify critical incidents and record the positive of negative aspects of their use of the application.

A decade later, Petrie *et al*. identified several other dimensions that can be used to further categorize the remote evaluation activities [8]. Whether or not the participant is independent in evaluations or dependent on the presence of an evaluator, if communication is synchronous or asynchronous and whether the participant requires training in the tasks before conducting them are all aspects that need to be considered when planning remote evaluations.

Further to this, Andraesen *et al.* [1] indexed existing work on remote evaluation by the methodology followed, by the type of data collected and on the dimension of synchronicity. Within this analysis, there are examples of traditional asynchronous studies, such as diary studies [4], self-administered questionnaires and workflow logging [9] among others. These methods are often supported by often general-purpose tools that are specialized for purposes of the evaluation or through bespoke implementations. While these bespoke tools are very useful, the specialization of the tool to the evaluation makes it potentially difficult to reuse components from these implementations in future remote evaluation protocols.

Due to the wide variety of techniques available, and the different types of information that can be collected, it is perhaps unsurprising that no unified framework or tool support has emerged to support the evaluator in the tasks of designing, deploying and conducting remote evaluations with users. In the following sections the architecture and prototype implementation of such a tool suite for conducting evaluations on websites are presented.

## 3 Requirements for Managing Evaluations

The phases of remote accessibility or usability evaluation are similar to those in standard co-operative evaluation techniques [7]. Evaluators must recruit a representative group of participants and record their demographic information for use in later analysis. They must design experimental tasks that are representative of what the users will do in the application. They must deploy the evaluation to the users, usually with some instruction of how to perform the task. Finally, the data must be analyzed through a variety of qualitative and quantitative methods. Currently, only the final step of this process is supported through automated tools statistical for quantitative analysis or qualitative analysis tools.

For the first three stages of the process, the authors analyzed existing literature and examined five investigations for requirements regarding what evaluators and participants would require from a suite of tools supporting remote evaluation. These five investigations were conducted under the auspices of the Benchmarking Tools and Methods for the Web (BenToWeb) project[1]. This project had the goal of producing tools and methods that aided in the validation and evaluation of websites for accessibility. Within that project there were several development efforts in which data was

[1] www.bentoweb.org retrieved 03/2009

required to inform the design of new tools. Among the investigations conducted by the project were: language simplification, navigational consistency, perception of colour contrast and colour confusion zones for people with colour vision deficiencies. Each of these investigations used a combination of remote tools, such as surveys/questionnaires and bespoke applications, to collect information from users throughout Europe. In addition to these investigations, there was an initiative to create a test suite for checking new accessibility tools as they come on the market for correctness and completeness. This initiative also collected information from remote users with disabilities about the success or failure of web implementations [10].[2] After an analysis of the functionality that was used in all of these remote data collection activities, the following sets of functional requirements were defined for a general online testing framework.

For the recruitment and registration of users, the test suite must have the following available:

- Participants must be able to record their demographic information such as age, sex, functional disability information and nationality for analysis purposes.
- Private aspects of participants' information, such as their names, addresses or billing information must not be associated with their evaluation result data.
- Participants must be able to specify technology configurations and their experience with different types of technology. This record must include: general operating systems, user agents and assistive technology to form a personal profile under which they will conduct trials.
- Many such technology profiles may be required for each participant to account for different contexts of use (e.g. home versus work).
- Participants must be able to select which and how many trials they would like to participate in from the overall set of remote trials available.
- Participants must be able to specify whether they would like to have their direct actions recorded for analysis, as opposed to reporting critical incidents or completing surveys. Some users may be uncomfortable with such monitoring components, or the components may conflict with aspects of their technology.
- A record of the trials completed by participants must be kept accurately so that participants can be appropriately reimbursed for their activities.

For the evaluator, the key user requirements come from the need for flexibility in specifying the applications that will be tested as well as the types of information that will be requested from the participants. These requirements include:

- Evaluators must be able to specify both custom built websites and websites in "the wild" for purposes of evaluation.
- Evaluators must be able to specify tasks on a website that are to be completed by a subgroup of users. These tasks are referred to as trials hereafter.
- Evaluators must be able to specify an arbitrarily large set of questions to ask the user before or after a trial, or set of trials, has been completed.

[2] This work resulted in two tools, Parsifal and Amfortas, that have provided inspiration for the test suite discussed in this paper. Thus the name of the test suite, Klingsor is drawn from the same source as those, the opera *Parsifal* by Wagner.

- Evaluators must be able to specify alternate choice, multiple choice, Lickert scale or open answer questions for each trial.
- Evaluators must be able to specify subgroups of the user population for engagement in remote trials. This can include grouping people by user characteristics such as technology experience, or through specification of the availability of technology to a participant (e.g. trials intended for screen reader users).
- User group profiles may be reused between remote evaluations. As such, evaluators must be able to save profiles about subgroups.
- Evaluators must be able to provide instructions and training documents through the test suite for each trial or set of trials.
- Evaluators must be able to specify an external tool that can be used for remote monitoring of user activities.
- Evaluators must be able to retrieve extracts of collected data from the system at any time in order to perform analyses on the data.

The system itself has the core requirement that it must perform correct profile matching between the trials specified by the evaluator and the participants. When complete, different users will be presented with different subsets of trials that are appropriate to their personal preferences, their technology and their experience.

## 4 Klingsor: A Remote Evaluation Tool Suite

With the above requirements in place, the authors proceeded to design and implement a prototype tool suite to manage remote evaluations of websites. In this section the architecture and implementation of these prototypes are discussed.

### 4.1 Architecture

The overall architecture of Klingsor is presented in Figure 1. The evaluation suite has two distinct tools, one for the evaluator and one for the participants, each of which have interface components connected to processing components that read and store different types of data regarding trials and participant respectively. A third processing component performs matching between the data models for the trials and profiles of participants.

In the evaluator interface, the evaluator is able to edit information regarding the evaluation or individual evaluation trials. For the evaluation, briefing and debriefing information can be presented to the participants at appropriate times, and instructions about the website/web application being used in the evaluation can be included to aid in training them. For each trial, the evaluator can specify the following pieces of information:

- *Target users*: this includes functional disability information and other information about the desired participants.
- *Target technology*: a description of the types of technology that are needed in the trial. For example, a screen reader or other assistive technology may be required for a particular trial. When this information is entered, the target user information can be updated with user experience on that particular technology.

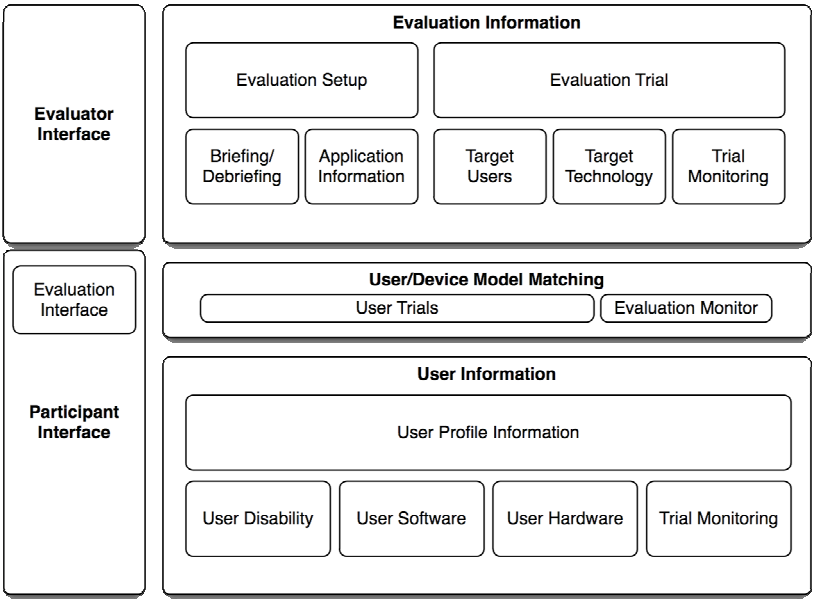


**Fig. 1.** The architecture for the Klingsor evaluation suite

- *Trial monitoring*: an indication as to whether a trial requires that the user be monitored by an external application or augmentation to the browser. This can include a reference to the monitoring software for installation on the participants' computers.

In the participants' interface, each participant can enter his or her own personal information as well as several technology profiles under which they will undertake the trials. Only one profile can be marked as active at any given time. This active profile will be the one used for matching with specific trials.

When a user logs into the participant interface, the model-matching component will compare the active profile of the user with the information regarding the trials. When there are trials available that match the participant's profile, he/she will be presented with a list of evaluation trials he/she may undertake. If the user undertakes any of the displayed trials, he/she will be credited with the completion for purposes of recompense.

### 4.2 Implementation

The Klingsor prototype evaluation suite has been implemented to be a cross-platform Java servlet application with Java Server Pages (JSP) serving as the web interface for the user. The JSP code has been engineered to produce static code that is rendered in the users' web browsers (either evaluators or participants) to avoid accessibility issues that may arise from dynamic content (such as enriched internet applications implemented in AJAX). As a result, the web application relies on heavy use of form fill-in

interactions with the user, leading them through the process of specifying evaluations (for the evaluator) or accessing currently available trials (for the participant).

For data storage, the components are implemented on a mySQL database that has been decoupled from the application code. The intention is that the data aspect of the application can be replaced with a Resource Description Framework (RDF) data repository or other data modeling language should the need arise.

## 5 Future work

The prototype implementation of the Klingsor suite is in the final stages of verification and validation. When complete, the tool will be deployed for use by student evaluators at both the University of York and the Technical Universität Dresden for an initial pilot. This work is scheduled for autumn 2009.

## 6 Conclusions

This paper has presented the requirements and architecture for a suite of tools for evaluators to manage and deploy evaluation protocols to remote users. Such a tool suite must be flexible enough to account for the variety of users that an evaluator may wish to engage, as well as in what types of information will be collected from participants.

This architecture collects information regarding the preferences and technology configurations of the participants and uses it to match them with evaluation trials specified by an evaluator.

A prototype implementation is being prepared for a large-scale deployment in the coming months in which evaluators will test the functionality for its usability, its accessibility and how fit-to-purpose the tool suite is for their needs. When complete, this tool will provide a new, innovative way to manage and conduct remote evaluations in both research and practice communities.

## References

1. Andreasen, M.S., Nielsen, H.V., Schrøder, S.O., Stage, J.: What happened to remote usability testing?: an empirical study of three methods. In: Proceedings of the SIGCHI Conference on Human factors in Computing Systems (2007)
2. Castillo, J.C., Hartson, H.R., Hix, D.: Remote usability evaluation: can users report their own critical incidents. In: CHI 1998: Proceedings of the SIGCHI Conference on Human factors in Computing Systems, pp. 253–254 (1998)
3. Dray, S., Siegel, D.: Remote possibilities?: International usability testing at a distance. Interactions 11(2), 10–17 (2004)
4. Hartson, H.R., Castillo, J.C.: Remote evaluation for post-deployment usability improvement. In: AVI 1998: Proceedings of the working conference on Advanced visual interfaces, pp. 22–29. ACM, New York (1998)

5. Hartson, H.R., Castillo, J.C., Kelso, J., Neale, W.C.: Remote evaluation: the network as an extension of the usability laboratory. In: CHI 1996: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 228–235. ACM, New York (1996)
6. Hill, W.C., Terveen, L.G.: Involving remote users in continuous design of web content. In: Proceedings of the conference on Designing interactive systems: processes, practices, methods, and techniques, pp. 137–145 (1997)
7. Monk, A., Wright, P., Haber, J., Davenport, L.: Improving your human-computer interface: A practical technique. Prentice Hall International (UK) Ltd., Englewood Cliffs (1993)
8. Petrie, H., Hamilton, F., King, N., Pavan, P.: Remote usability evaluations with disabled people. In: Proceedings of the SIGCHI conference on Human Factors in computing systems (2006)
9. Siochi, A.C., Ehrich, R.W.: Computer analysis of user interfaces based on repetition in transcripts of user sessions. ACM Trans. Inf. Syst. 9(4), 309–335 (1991)
10. Strobbe, C., Koch, J., Vlachogiannis, E., Ruemer, R., Velasco, C.A., Engelen, J.: The BenToWeb test case suites for the web content accessibility guidelines (WCAG) 2.0. In: Miesenberger, K., Klaus, J., Zagler, W.L., Karshmer, A.I. (eds.) ICCHP 2008. LNCS, vol. 5105, pp. 402–409. Springer, Heidelberg (2008)
11. Winckler, M.A.A., Freitas, C.M.D.S., de Lima, J.V.: Usability remote evaluation for www. In: CHI 2000: CHI 2000 extended abstracts on Human factors in computing systems, pp. 131–132. ACM, New York (2000)