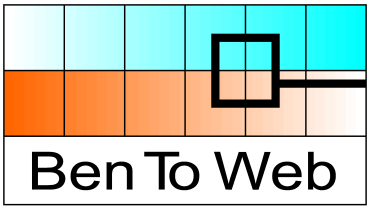
Benchmarking Tools and Methods for the Web (FP6—004275)



Sixth Framework Programme
Information Society Technologies Priority

D3.8 Report on final evaluation framework for project monitoring

Contractual Date of Delivery to the EC:	30 September 2007 + 45 days
Actual Date of Delivery to the EC:	29 November 2007
Editor:	Helen Petrie (York)
Contributors:	Helen Petrie, Christopher Power (York)
Workpackage:	3
Security:	Public
Nature:	Report
Version:	B
Total number of pages:	16

Keywords: Web accessibility, Web Content Accessibility Guidelines 2.0 (WCAG 2.0), test cases, evaluation methodologies, assistive technologies, disabled users.

DOCUMENT HISTORY			
Version	**Version date**	**Responsible**	**Description**
A	31 August 2007	York	First draft.
B	15 November 2007	York	Final version.

Table of Contents

List of Figures

1 Executive Summary

The BenToWeb Project has successfully used a wide range of different validation and evaluation methodologies, as appropriate to the different tasks as situations.

It has used standard methodologies:

- interviews: were used in the elicitation of information from web developers and website commissioners and owners;
- laboratory-based studies: were used in the study of navigational consistency parameters and the perception of colour combinations by people with colour vision deficiencies.

It has used modern twists on standard methodologies:

- online surveys: were used in the elicitation of information from web developers and website commissioners and owners;
- the Cello expert inspection evaluation method: was used to investigate the usability of existing accessibility evaluation tools and the BenToWeb navigation consistency module.

Finally, and perhaps most importantly the project has innovated in the area of validation and evaluation methodologies, creating an expert methodology for the validation of test cases, supported by the Parsifal tool and a large scale remote user testing methodology for the user testing of unresolved components of the test cases, supported by the Amfortas testing environment.

2 Introduction

The BenToWeb Project has used a wide range of validation and evaluation methodologies in the course of the project. The methodologies chosen have been appropriate to the stage of the project and the type of validation or evaluation task to be undertaken.

This deliverable will provide an overview of the validation and evaluation methodologies used and comment on some of the more innovative methodologies.

3 Overview of Validation and Evaluation Methodologies Used in the Project

3.1 User requirements and characteristics elicitation: online surveys and interviews

At the beginning of the project, investigations into the levels of interest and knowledge in web accessibility of both web developers and website owners and commissioners were undertaken. For these investigations, an online survey and follow-up phone interviews were undertaken, two very standard methodologies for the investigation of user requirements and characteristics. The results of these investigations were reported in Deliverable 3.3.

3.2 Expert evaluations

To evaluate the strengths and weaknesses and usability of existing accessibility evaluation tools, an expert evaluation methodology was used. This evaluation might have been conducted as a user study, asking web developers to undertake accessibility evaluations of their own web sites or of specially constructed web sites. However, two factors argued against this approach. Firstly, this would have been a very big commitment by real web developers of their time, which would have needed to include some training in accessibility issues and in how to use the accessibility evaluation tools. Secondly, it was known that the functionality and usability of many of the tools to be evaluated is quite poor, so it was not deemed to be a good use of professional web developers time to ask them to undertake such an evaluation. A study could have been done with students, but this was felt to lack ecological validity as a user study and would probably not have elicited as much useful information as an expert evaluation.

A standard usability inspection evaluation could have been undertaken, asking a number of experts to go through each of the accessibility evaluation tools, then asking them to come together and discuss their observations and agree on ratings of the usability problems they had found (Nielsen, 1993). However, each of the accessibility evaluation tools to be evaluated requires a certain amount of learning, which is often not simple, as the tools provide little by way of online help or manuals. Several members of the BenToWeb team had used particular tools, so it

was decided to use a variation of the standard usability inspection method, and have a good of experts work together to evaluate the tools. This method has been developed in the European funded Emmus (European Multimedia Usability Services) Project[1] and is known as the Cello method.[2] In this variation, the experts work through a system together, discussing potential usability problems, but then identify and rate the problems privately, so a different number of problems and different ratings are obtained from each expert. The results of the Cello evaluation of existing accessibility evaluation tool are reported in Deliverable 3.2.

The BenToWeb team felt this was a very successful expert evaluation method and it was used again in the evaluation of the usability of BenToWeb navigation module, which was reported in Deliverable 3.6. The York team were so impressed by this method that they are planning a study to compare the effectiveness (i.e., the proportion of usability problems detected) and efficiency (i.e., total person power required) of the Cello method in comparison with the standard expert usability inspection method and user testing for the evaluation of web sites and applications.

3.3 Laboratory studies with users

A number of classic laboratory based studies with users were also undertaken as part of the BenToWeb project. These studies were appropriate when detailed information was required from users in controlled settings. For example, the investigation of the validity of the colour vision deficiency module needed to be undertaken with participants with different colour vision deficiencies. This study could have been undertaken in a variety of situations, but because of the difficulty of recruiting participants with different colour vision deficiencies, it was easiest to ask them to come to the lab to undertake the testing. The results of this study are presented in Deliverable 3.5.

Two studies were also undertaken of some of the components in the consistency of website navigation that affect users. These were undertaken in order to provide parameters for the BenToWeb navigational consistency module. Again, these studies might have been undertaken in a variety of situations, but it was decided to run the studies in the laboratory as this allowed the research team to record the participants detailed interaction with the websites. This proved very important, as differences were found between the participants' perception of the consistency of website navigation components and how it actually affected their performance. This would not have been detected if the studies had

[1] http://www.ucc.ie/hfrg/emmus/
[2] http://www.ucc.ie/hfrg/emmus/methods/cello.html

been undertaken online, as originally planned. The results are presented in Deliverable 3.6.

3.4 Online studies

In other situations, an online study proved very useful to gather information from large number of participants in their natural web browsing environments. For example, a large study of the parameters of colour contrast between text and background was conducted as an online study. 180 people each evaluated 54 sentences written in different coloured text on different coloured backgrounds for ease of reading. Although the effort to create the website and backend for such a study should not be underestimated, it provided an excellent vehicle for collecting the relevant information. The results of this study are presented in Deliverable 3.4.

3.5 Innovative methodologies in BenToWeb

In two areas, particularly innovative methodologies were employed in the BenToWeb Project, which required development of new techniques and new programs to support them. These were the validation methodology for the validation of test cases for the WCAG2 techniques and the large scale remote user evaluation of scenarios for the test cases. These will be discussed in more detail in the next sections.

4 The BenToWeb Methodologies: Test Case Validation and large scale remote user testing

4.1 Validation of test cases for WCAG2 drafts

One of the tasks in the BenToWeb Project was to create test suites of test cases to support testing of automatic testing tools and manual testing methodologies for conformance to the Web Content Accessibility Guidelines (WCAG). During the lifetime of the project, a number of drafts of Version 2 of WCAG were produced, so three test suites were produced and validated.

To validate the test cases an interactive tool was developed within the BenToWeb Project, Parsifal. Each test case contains the following information:

- one or more test files containing an web page fragment implementing the relevant WCAG2 technique. The technique can either be applied correctly, or with one error that might be made by web developers in its application;
- a description of what the test file(s) contains;
- a description of the purpose of the test file(s);
- a proposal for what test mode will be appropriate for the test case – automatic, one expert, or several experts; and
- a prediction as to whether the content of the test file(s) will be accessible or inaccessible to users.

After each test case was initially proposed by a test case author, it was validated in a two stage process by two accessibility experts. If the outcome of the test case was unclear, even when discussed by the two accessibility expert validators and the test case author, a set of scenarios were created for user testing, to establish outcomes with appropriate user groups.

This validation process is presented in Deliverables 3.7a and 3.7b. It ensures that the test cases are informed by the knowledge of several accessibility experts in a structured manner and where appropriate also

informed by information from user testing. The process for conducting the large scale user testing required is described in the next section.

4.2 Large scale remote user testing to support test case validation

Conducting adequate evaluations of emerging systems with real users in naturalistic, ecologically valid situations of use has always been a great difficulty for HCI researchers and practitioners. All too often delays in the technical development and implementation, lack of resources, lack of access to users and their environments or just general lack of will means that systems fail to get evaluated or get evaluated inadequately. This results in systems that are difficult to use, are under-used or not used at all and creates skepticism about the effectiveness of HCI as an enterprise.

This situation may well be getting worse with the spread of computing systems into everyday life and the complexity of those systems increases. Whereas until recently evaluations needed to take place only on a desktop platform, now evaluations may need to consider multiple platforms such as home and mobile devices, as well as office devices. Within a single platform, multiple operating systems (OSs) may need to be tested, and using multiple platforms almost inevitably involves multiple OSs. Closely related to the range of device platforms is the increasing range of situations of use – no longer only the office, but also the home (including different situations in the home) and increasingly public and outdoor spaces.

Another very important factor is the increasing range of users, not only increasingly technologically sophisticated users (although it is important to create systems that are usable and satisfying to these users), but also less technologically sophisticated users, as well as older and disabled users.

These latter two groups are important to include in evaluations for different reasons. Older users are important because the first generation of users of PCs in now aging (Bill Gates and Steve Jobs are both now over 50), but will undoubtedly want to continue to use the full range of technologies they have become used to, and may increasingly find more difficult to use. People who are already considerably older are becoming more interested in technology, as it helps them overcome the problems of aging and stay in touch with younger family members and friends. This increasing openness to technology is combined with the increasing aging of the population. In many developed regions of the world, the number of people over 60 has already exceeded the number of children (aged under 15 years) (United Nations, 2007). This aging of the population will increase for at least the next 40 years, with the percentage of the

population over 60 years, currently 11%, predicted to rise to 22% by the year 2050 (United Nations, 2007). Thus older users of computing systems will become an ever more important user group for evaluations.

An additional problem for such evaluations is that older users are more heterogeneous in their responses than younger users (Newell and Gregor, 2002) so a larger number of participants is required in an evaluation to provide a specific level of power for statistical comparisons (Howell, 2007). Users with disabilities are increasingly important as technological developments have provided access to computing systems for many people with disabilities. Blind people now use screen readers and Braille displays (Mynatt and Weber, 1996; Petrie, O'Neill, and Colwell, 2002; Vidal-Verdu and Hafez, 2007), partially sighted people use screen magnification programs (Blenkhorn, Evans, King, Kurniawan and Sutcliffe, 2003), people with physical disabilities use alternative input systems such as sip and puff or infra-red control systems (Cook and Hussey, 2002). All these assistive technologies add considerable complexity to evaluations – mainstream systems need to work with assistive technologies, but the average usability laboratory does not have a range of these systems (let alone all the brands and versions of them) and the average usability expert is not familiar with how they all work.

Partly due to this increasing complexity of the evaluation landscape, there has been a recent renewal of interest in remote evaluation techniques, following interest in this topic in the mid 1990s (Hartson, Castillo, Kelso and Neale, 1996). Petrie, Hamilton, King and Pavan (2006), building on the earlier work from Hartson et al (1996), classified remote evaluations on a number of dimensions to facilitate comparisons. For example, evaluations are synchronous if participant and evaluator need to participate at the same time, albeit in different locations. Remote evaluations can also vary in the extent to which participants are independent from the evaluator, and require training in order to undertake the evaluation.

Brush, Ames and Davis (2004) compared synchronous local and remote evaluations, and found that the number of usability problems elicited, their type and severity, were similar between the two techniques. Petrie, Hamilton, King and Pavan (2006) found that while quantitative data were similar between asynchronous local and remote evaluations, the amount and richness of qualitative data was less for remote evaluations. Andreasen, Nielsen, Schrøder and Stage (2007) also found that synchronous remote evaluation was very close to a standard local evaluation, but that asynchronous remote evaluations produce fewer usability problems. However, these remote evaluation techniques are still based around standard user evaluation methodologies which involve having a small number of users work through a small number of tasks

with a system. This means it is very time-consuming to undertake an evaluation with the variety of users, situations of use and devices required for a comprehensive evaluation.

For example, if one wanted to undertake an evaluation of particular aspects of interaction with websites (for example, the usability of different navigational schema, as was undertaken in the BenToWeb Project on a small scale) on three different platforms with three different user groups each using at least two combinations of browser and/or assistive technology, with a minimum of five users per combination (to ensure valid and robust results for the evaluation), one would need at least 90 users. It is therefore not surprising that such comprehensive evaluations are relatively rare.

The BenToWeb Project developed a new methodology for remote web-mediated large scale evaluations that addresses these issues. Instead of asking remote participants to undertake standard multi-step tasks for an evaluation, for which synchronization with the evaluator is needed to ensure the same level of results as can be achieved with a local evaluation, the methodology presents remote participants with very small scenarios which test one very specific aspect of web accessibility. By using many such scenarios with many users (far easier to manage in the remote evaluation situation), it becomes much more feasible to undertake comprehensive evaluations of considerable complexity.

The use of very small scenarios may not be suitable for all system domains, but within the BenToWeb Project it was found to be extremely useful in the area of web accessibility. However, we believe it will have application in many other areas. The fact that scenarios are very small and only require one or two answers (often a quantitative measure and a qualitative measure) also means that participants will be more motivated to complete their task and produce both good quality quantitative and qualitative data.

The steps of the remote user testing methodology are outlined in Figure 1, below. The methodology starts with the test cases which contain the test files and purpose of the test case. The test cases are used to develop scenarios to present the test material to users and generate responses from them. The scenarios specify which types of users, with which technologies should evaluate any particular test case. A web-mediated system, AMFORTAS, is then used to match users to scenarios, so that the scenarios can be evaluated by appropriate users in appropriate naturalistic settings of use.

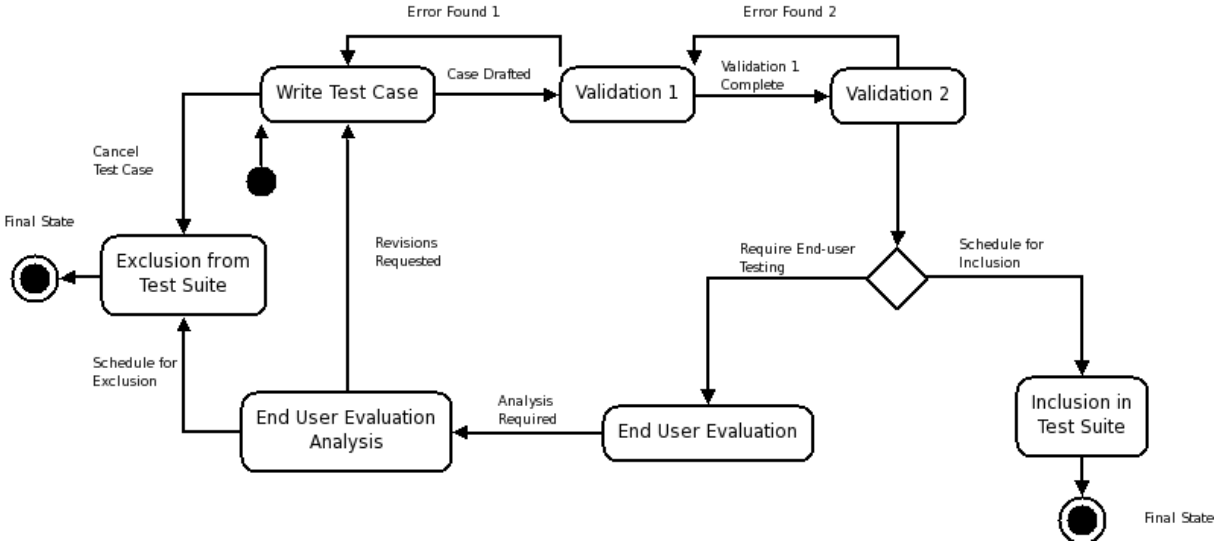


Figure 1: Test Case Validation Process.

The validation methodology for test cases was refined over the three iterations of the testing. The most detailed discussion of the remote user testing methodology can be found in Deliverable 3.7b.

5 Discussion

The BenToWeb Project has successfully used a wide range of different validation and evaluation methodologies, as appropriate to the different tasks as situations. It has used standard methodologies, such as interviews and laboratory-based studies; it has used modern twists on standard methodologies, such as online surveys and the Cello expert inspection evaluation method. Finally, and perhaps most importantly the project has innovated in the area of validation and evaluation methodologies, creating an expert methodology for the validation of test cases, supported by the Parsifal tool and a large scale remote user testing methodology for the user testing of unresolved components of the test cases, supported by the Amfortas testing environment.

The use of these methodologies have raised some interesting questions and possibilities, which are being explored further outside the scope of the BenToWeb Project.

6 References

Andreasen, M. S., Nielsen, H.V., Schrøder, S.O. and Stage, J. (2007). What happened to remote usability testing?: an empirical study of three methods. Proc. CHI 2007. New York: ACM Press.

BenToWeb Project (2006). Deliverable 3.2: Report on usability of existing Web accessibility E&R tools.

BenToWeb Project (2006). Deliverable 3.3: Report on the survey of Web site designers and commissioners of Web sites.

BenToWeb Project (2007). Deliverable 3.4: Evaluation of colour-deficiency module.

BenToWeb Project (2007). Deliverable 3.6: Evaluation of colour-contrast module.

BenToWeb Project (2006). Deliverable 3.7a: Evaluation and validation reports for test suite: (X)HTML and CSS2.

BenToWeb Project (2007). Deliverable 3.7b: Evaluation and validation reports for test suite: (X)HTML and CSS2 (2nd update).

Blenkhorn, P.; Evans, G., King, A.; Kurniawan, S.H., and Sutcliffe, A. (2003). Screen magnifiers: evolution and evaluation. IEEE Computer Graphics and Applications, 23(5), 54–61.

Brush, A.J., Ames, M. and Davis, J. (2004). A comparison of synchronous remote and local usability studies for an expert interface. Proc. CHI 2004. New York: ACM Press.

Cook, A.M. and Hussey, S. (2002). Assistive technologies: principles and practice (2nd edition). Philadephia: Elsevier Science.

Hartson, H.R., Castillo, J.C., Kelso, J., and Neale, W.C. (1996). Remote evaluation: the network as an extension of the Usability Laboratory. Proc. CHI 1996, New York: ACM Press.

Herramhof, S., Petrie, H., Strobbe, C., Vlachogiannis, E., Weimann, K., Weber, G., and Velasco, C. A. (2006). Test case management tools for accessibility testing. In: Miesenberger K. et al (Eds.), Proceedings of the 10th International Conference ICCHP 2006 (Linz, Austria, July 2006), LNCS 4061, Berlin-Heidelberg: Springer-Verlag.

Howell, D.C. (2007). Fundamental statistics for the behavioral sciences (6th edition). Wadsworth.

International Standards Organization. Ergonomics of Human System Interaction, Part 11: Guidance on usability. Geneva: International Standards Organization.

Mynatt, B. and Weber, G. (1996). Nonvisual Presentation of graphical user interfaces: contrasting two approaches, Proc. CHI 1994. New York: ACM Press, 166-172.

Newell, A. F. and Gregor, P. (2002). Design for older and disabled people – where do we go from here? Universal Access to the Information Society, 2, 3 – 7.

Nielsen, J. (1993). *Usability engineering*. San Diego, CA: Academic Press.

Petrie, H., Hamilton, F., King, N., and Pavan, P. (2006). Remote usability evaluations with disabled people. Proc. CHI 2006. New York: ACM Press.

Petrie, H., O'Neill, A-M. and Colwell, C. (2002). Computer access by visually impaired people. In A. Kent and J.G. Williams (Eds.), Encyclopedia of Microcomputers Volume 28. New York: Marcel Dekker. 1

Strobbe, C., Engelen, J., Koch, J., Velasco, C., Vlachogiannis, E. and Ortner, D. (2007). The BenToWeb XHTML 1.0 test suite for the Web Content Accessibility Guidelines 2.0 - Last Call Working Draft. In C. Stephanidis (Ed.), Universal Access in Human-Computer Interaction. Applications and Services, Proc. HCI International, (22.-29.7.2007 Beijing), LNCS 4556, Berlin-Heidelberg: Springer-Verlag.

United Nations (2007). World Population Ageing 2007. New York: United Nations. Available at: http://www.un.org/esa/population/publications/WPA2007/wpp2007.htm

Vidal-Verdu, F. and Hafez, M. (2007). Graphical tactile displays for visually impaired people. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 15(1), 119 – 130.

Web Accessibility Initiative. http://www.w3.org/WAI/

Web Accessibility Initiative. (2006). Web Contents Accessibility Guidelines. Last Call Working Draft (27 April 2006). Available at: http://www.w3.org/TR/2006/WD-WCAG20-20060427/