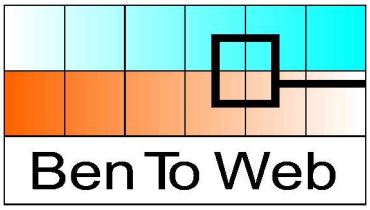
Benchmarking Tools and Methods for the Web (FP6—004275)



Sixth Framework Programme
Information Society Technologies Priority

D3.1 Report on evaluation framework for project monitoring

Contractual Date of Delivery to the EC:	28th February 2005
Actual Date of Delivery to the EC:	31st March 2005
Editor:	Helen Petrie (City)
Contributors:	Helen Petrie (City), Christophe Strobbe (KULRD), Gerhard Weber (MMC), Carlos A Velasco (FIT)
Workpackage:	3
Security:	Public
Nature:	Report
Version:	E
Total number of pages:	24

Keywords: Web accessibility, evaluation and repair tools, evaluation methods, user testing, monitoring.

Table of Contents

List of Tables

1 Executive Summary

This deliverable will outline how we will monitor progress in different parts of the BenToWeb Project and evaluate the outputs of the different components of the project.

2 Introduction

This deliverable will outline how we will monitor progress in different parts of the BenToWeb Project and evaluate the outputs of the different components of the project. After a detailed analysis, we have identified the following components:

- monitoring dissemination through the project's Web site and linked activities such as the WAB cluster and the ENABLED Project (see section 3, below). For exploitation issues, please refer to deliverable D2.1;
- the User Panel (see section 4, below);
- test suites (see section 5, below);
- expert and user evaluations of project outcomes (see section 6, below); and
- evaluation and benchmarking of existing accessibility Evaluation and Repair tools (see section 7, below).

3 Monitoring dissemination and exploitation

A major tool for monitoring dissemination will be the BenToWeb Project Web site.[1] Any new public documents from the project will be listed and available for download on the appropriate page of the Web site with bibliographic data. The Web site already announces the project deliverables that will become available there.

Monitoring of publications through our Web site will include checking for accessibility, as BenToWeb aims to inform user organisations as well as researchers and any other interested parties. All deliverables are prepared using the non-proprietary formats of OpenOffice[2] in order to support the translation process to other formats like XHTML.

The second tool for monitoring WAB cluster activities is its Web site.[3] Links to some deliverables such as the different surveys will be established and checked for completeness together with all updates on the Web sites. This includes also validation of links to the ENABLEDWEB integrated project,[4] and in particular with its Accessible Web Contents activity (AWC).

The monitoring of the exploitation plans (see D2.1) will:

- follow technical achievements according to the work plan;
- include monitoring of developments of other approaches and methods to evaluate accessibility; and
- review newly founded national projects and initiatives on accessibility such as the eAccessibility public consultation.[5]

It is the policy of WAI to be pre-competitive in its development of recommendations. As no commercial interest seems to grow for developing separate accessibility guidelines, this situation will allow to contribute to new recommendations through the participation in WAI groups and at the same time develop our tools and consultancy services. Monitoring therefore includes mostly monitoring of WAI recommendations.

As D2.1 "Plan for using and disseminating knowledge" describes various exploitation activities, monitoring of these is part of work package

[1] http://bentoweb.org/
[2] http://www.openoffice.org/
[3] http://www.wabcluster.org/
[4] http://www.enabledweb.org/
[5] http://europa.eu.int/information_society/policy/accessibility/com_ea_2005/a_documents/com_consult_res.html

management activities and will be documented in the corresponding management reports.

4 Monitoring the User Panel

Initial discussions within the BenToWeb Project revealed that a number of User Panels will be required for the work of the project, not just one. It is envisaged that the following User Panels will be required:

- **Main User Panel (disabled people):** people who are blind, partially sighted (including people with colour vision deficiencies), physically disabled, deaf, hard of hearing, dyslexic and cognitively disabled. This panel should also include people with a variety of levels of experience on computers and the WWW, as well as people who use a variety of assistive technologies, with different levels of expertise.
- **Older User Panel:** a panel of people aged 65 and over, with the typical ranges of disabilities experienced by this population. As with the Main User Panel, this panel should include people with a variety of levels of experience on computers and the WWW, as well as people who use a variety of assistive technologies, with different levels of expertise.
- **Colour Vision Deficiencies Panel:** a panel of people with the full range of colour vision deficiencies. This panel will not need to be a numerically representative sample of colour vision deficiencies, but we will need people with all the different possibility colour vision configurations, so we can call on them to test Web pages with different colour combinations and contrasts.
- **Web developers:** A panel of Web developers will need to be recruited for the evaluation of existing Web accessibility E&R tools (in Task 3.4 of WP3), and the survey (in Task 3.6 of WP3). It will be quite difficult to recruit a representative sample of Web developers to give up the substantial amounts of time required to undertake evaluations of E&R tools. Those likely to be willing to do so are undoubtedly those who are interested in accessibility, and while relevant, this is not the population we are really interested in, that being those Web developers who are ignorant and apathetic about Web accessibility. It should be therefore easier to recruit web developers to take part in a survey, which will only take a short time to undertake.
- **Web commissioners/owners:** a panel of Web commissioners/owners will need to be recruited to undertake for the survey in Task 3.6 of WP3). It will also be difficult to recruit the types of Web commissioners/owners of most interest to the project: those who are ignorant and apathetic towards Web

accessibility. However, the project will undertake major publicity efforts to reach these groups of people.

A number of activities are underway to form the user panels:

- A recruitment questionnaire has been developed. This will be used by partner organizations who have strong links with the relevant types of people required for the User Panels (e.g., University of the Aegean, Accessibility, City University and ISdAC) to directly recruit people.
- A recruitment page for the Web site (yet to go public).
- Publicising the survey to web developers and commissioners/owners: a number of partner organizations have publicised the survey in different ways (e.g., having articles in newsletters, placing links on websites).
- Creating links to other User Panels: to create the Colour Deficiencies User Panel, City University is liaising with the Department of Optometry and Visual Science, who have a panel of people with colour deficiencies, and the International Achromatopsia Network, to recruit sufficient numbers of people with the rarer colour vision deficiencies.

To ensure that the make-up of the User Panels meets the criterion set out above, BenToWeb partners City University and ISdAC will monitor the make-up of each User Panel and where appropriate, make efforts to correct the composition of each Panel.

5 Monitoring the Development and Use of the Test Suites

This chapter can be considered as the QA Process Document (QAPD) for the development and use of the test suites. Producing a QAPD is required by guideline 4 ("Define the QA process") of the Operational Guidelines[6] of the W3C QA Framework (which is, admittedly, only work in progress).[7]

5.1 Objectives of the Test Suites

The test suites for the Web Content Accessibility Guidelines 2.0 have two objectives:

1. provide a basis for benchmarking of evaluation and repair (E&R) tools for web accessibility, and
2. provide a set of sample files that web developers can refer to as illustrations of technology-specific techniques.

For BenToWeb, the first objective is far more important than the second, but the WAI GL Working Group uses test files only to fulfil the second objective.

Benchmarking E&R tools is outside the scope of this work; the test suites will be provided for help and information only. However, we intend to produce as comprehensive test suites as possible.

The test suites could be hosted at W3C once developed; in that case, the W3C Document and License[8] will apply. Also, after development they could be mirrored at various sites in order to simplify access and enhance availability to the community. BenToWeb intends to host a CVS server to facilitate the development of the test suites.

[6] http://www.w3.org/TR/2003/CR-qaframe-ops-20030922/.
[7] Other examples of QA Process Documents are the DOM Conformance Test Suites Process Document (http://www.w3.org/2002/01/DOMConformanceTS-Process-20020115), the W3C XML Schema Test Collection document (http://www.w3.org/2001/05/xmlschema-test-collection.html), the document "W3C SOAP Version 1.2 test collection - How to contribute" (http://www.w3.org/2000/xp/Group/1/10/ts-contribution) and the Conformance Test Process For WCAG 2.0 (http://www.w3.org/WAI/GL/WCAG20/tests/ctprocess.html).
[8] http://www.w3.org/Consortium/Legal/copyright-documents-19990405.

5.2 Description of the Test Suites

Each of the test suites will consist of a set of files, a machine-readable list of the test files (most probably a list of links in an HTML file, because evaluation tools probably do not support any other mechanism for locating files) and metadata for the status and accessibility of each of the test files.

The files in a test suite belong to one or more groups (but the first two groups are mutually exclusive):

- Most files in a test suites are single units addressing one checkpoint (WCAG 1) or success criterion (WCAG 2). Each file corresponds to a test case.
- Some files may address multiple checkpoints or success criteria (they may especially violate several checkpoints or success criteria), to allow more thorough test for evaluation tools. Each file is a complex test case.
- For some checkpoints or success criteria (e.g., regarding consistent navigation), it is necessary to create groups of files that – taken together – implement or violate a single checkpoint or success criterion. These groups of files are compound test cases.

5.3 Contributing and Testing Use Cases

When a test case is added to a test suite, at least the following metadata should be made available:

- which checkpoint or success criterion is being addressed,
- whether the test case implements or violates the checkpoint(s) or success criterion/criteria (if this can be determined with certainty),
- how the test case should be used, i.e., a description guiding the members of the user panel, and
- administrative metadata such as author, date and version.

During the development of the test suites, it may also be useful to add comments for other test suite developers. These comments can be removed when the complete test suite is made available to the public.

User testing should provide a definitive statement on the accessibility of a test case, i.e., whether the test case violates a checkpoint or success criterion or not. This statement can conflict with the statement provided by the developer of the test case. The developer should then recheck the

test case and, if necessary, get more detailed information from the user panel. It should then become clear if a test case is accessible or inaccessible for the reasons that the developer had in mind. For difficult issues, it may be worthwhile to request the opinion of external users or experts through mailing lists such as the WAI Interest Group mailing list.

User testing may not always lead to an unequivocal statement on the accessibility of a test case: a test case may be accessible with one user agent-assistive technology combination but not with another (possibly older) combination. For this reason it is important to record which user agent-assistive technology combination is used for the user test. This may prove to be especially important for complex test cases.

Each test case moves through a process with several steps. The W3C's Conformace Process for WCAG 2.0[9] shall be used. The steps in this process are the following (with adaptations for BenToWeb: final details depend on co-ordination with WAI):

- Test Case is received.
  Test Case is entered in BenToWeb's version control system (or W3C's Bugzilla), status is set to **unconfirmed**.
  Test Case is reviewed for completeness:
  - ✗ If not complete, it is returned to submitter with directions on how to complete.
  - ✗ If complete, go to next step.
- Test case is complete.
  Test Case status is changed to **new**.
- Test case is sent to mailing list for review (W3C) and reviewed by BenToWeb team / assigned to member of user panel (BenToWeb).
  Test Case status is changed to **assigned**.
- Test Case pending approval.
  After two weeks of being available for review on the mailing list (W3C) / X weeks of being available to the user panel (BenToWeb), status is changed to **pending**.
- Test Cases reviewed on Teleconference (W3C) / reviewed by user panel (BenToWeb).
  Test cases are reviewed on the WCAG Working Group teleconference call and then members vote in a straw poll to accept or reject the test (W3C) / Test cases are reviewed by the user panel and assigned an accessibility statement (BenToWeb)
  Test case is either accepted or rejected or holding:

[9] http://www.w3.org/WAI/GL/WCAG20/tests/ctprocess.html

- If accepted, status is changed to **accepted**
- If rejected, status is changed to **rejected**
- If unsure, status is changed to **holding**

The following table from the Conformance Test Process For WCAG 2.0 summarizes the types of status:

Status	Process
1 - unconfirmed	received and under review for completeness
2 - new	test case is complete
3 - assigned	under review on mailing list
4 - pending	review complete, waiting for decision on teleconference
5 - accepted	result of straw poll decision
6 - rejected	result of straw poll decision
7 - holding	unsure of status, maybe accepted or rejected

Note that test cases may be developed from scratch or borrowed from existing test suites, if license or copyright restrictions allow this. Deliverable D4.1 discusses relevant test suites that may be used as a source.

5.4 Communication between Developers and User Panel

Communication between the developers and the user panel will happen through evaluation reports for each of the test suites. These evaluation reports will be compiled after evaluation of the test suite by a user panel. The evaluation reports will be public documents. Follow-up discussions will happen primarily through e-mail, but other means (e.g., telephone) may also be used. These communications will not be public.

5.5 Timeframe

Test suite	Date
(X)HTML and CSS 2	August 2005
XForms	February 2006
SVG	October 2006
XHTML 2 and CSS 3	August 2007

XHTML 2 and CSS 3 are still under development and may not reach Recommendation status before 2007. This may affect the timing and coverage of the last test suite.

6 Framework for expert and user evaluations for monitoring project outcomes in WPs 4, 5, and 6

6.1 Introduction

The work undertaken in WPs 4, 5 and 6 of BenToWeb will produce much material of interest to end-users (i.e., disabled Web users) and a variety or other users, most particularly Web developers.

To ensure that appropriate user involvement in and monitoring of this work, a two stage framework has been proposed and is in the initial stages of evaluation.

6.2 Expert evaluations

The first stage of evaluation and monitoring for all work from within the BenToWeb project will be independent expert evaluations. By independent, we mean that the experts will be at different organizations from within the project and will not have been involved directly in the development work (although they may have had input into the specification process). This is particularly important so ensure that the evaluators do not have too close a knowledge of the module/outcome being evaluated.[10]

The precise nature of the expert evaluation will depend on the module/outcome being evaluated. At least three types of expert evaluation are already foreseen for the BenToWeb Project, although others may be required as the work develops:

- Expert evaluation against general usability principles, such as Heuristic Evaluation
  An example of such an expert evaluation has already started, in the work for Task 3.4 of WP3, "Evaluation of the usability of existing Web accessibility E&R tools." A preliminary report on the methodology for this type of evaluation is provided in Section 7 of this report.

[10] Cockton, G., Lavery, D. and Woolrych, A. (2003). Inspection-based evaluations. In J. Jacko and A. Sears (eds), The Human Computer Interaction Handbook. Mahwah, NJ: Lawrence Erlbaum Associates.

- Expert evaluation against Web accessibility guidelines, e.g., WCAG 1.0 (Chisholm et al, 1999) and 2.0 (Caldwell et al, 2004). Much of the work in WPs 4, 5 and 6 is aimed at providing better evaluation of parts of WCAG 1.0 and 2.0. Therefore the appropriate initial evaluations will be made by experts to see whether the work meets the criteria set out in these guidelines.
- Expert evaluation against more specific guidelines and other scientific criteria.
  However, the work will be attempting to elaborate on the guidelines set out in WCAG 1.0 and 2.0, so further expert evaluation will need to be made against more specific guidelines and other scientific criteria. For example, in the work on colour contrast, expert evaluations will look at the scientific evidence on contrast and determine whether the module supports the best and latest evidence on this topic.

6.3 User-based evaluations

When a module/outcome has been subject to expert evaluation, and is thought to be ready by the experts, the "gold standard" is always to subject it to user-based evaluations. Who the users are will depend on the nature of the module/outcome. It is envisaged that in the BenToWeb Project, two different types of user-based evaluations will be undertaken:

- Evaluations with disabled/older Web users - to ensure that pages approved by the E&R modules developed are accessible to these groups.
- Evaluations with Web developers to ensure that the E&R modules are usable by Web developers who do not have a detailed knowledge and commitment to Web accessibility, i.e., to typical Web developers.

For each of these types of user-based evaluation, the BenToWeb Project will develop a standard, but flexible, evaluation protocol that can be used in a variety of evaluation scenarios. These evaluation protocols will be developed and tested on emerging modules in the second 6 month period of the project.

6.3.1 Framework for user-based evaluations

Many of the specifics of the user-based evaluations will need to be developed and refined as particular evaluations are required. However, at the moment, we can outline the following framework for how face-to-face evaluations will be planned and undertaken. This framework applies both

to evaluations involving disabled participants and those involving web developers as participants.

In some circumstances, remote evaluations instead of face-to-face evaluations might be conducted. Great care needs to be taken to ensure that the data from remote evaluations are as rich and reliable as the data from face-to-face evaluations. In the web accessibility study conducted for the Disability Rights Commission[11], researchers at City University trained the participants in the methods for the remote evaluation in initial face-to-face sessions at the university. This allowed larger quantities of data to be collected than if the participants came to the university for all the sessions. This procedure might also be useful to the BenToWeb Project and has also been used by FIT and the University of the Aegean in the IRIS project (IST funded).

6.3.1.1 Step 1: Identify the purposes of the evaluation

May require some discussion and negotiation between developers and evaluators. It's always easy to assume as an evaluator that you understand why an evaluation is being undertaken. Some discussion at this point can reveal interesting unexplored areas on both sides (i.e., the developers and evaluators).

6.3.1.2 Step 2: Identify tasks and build scenarios

Evaluations always seem to work better if the participants are asked to undertake a realistic sequence of actions and tasks with a system or website, a "scenario of use". Even if large parts of this scenario of use have not be fully implemented, this gives the participants a far better sense of the overall system or website. However, the key aspect of this step is defining the tasks that the participants should undertake for the evaluation. These need to come from the purpose of the evaluation and need to be linked to Step 3, what is going to be measured. A typical evaluation involves six to eight key tasks to be undertaken with a system or website, but this can vary with the purpose of the evaluation. Considerable care needs to be taken that the tasks also make sense to the participants, and do not put too much stress on the participants. Tasks should be organized within the scenario from some tasks or components of tasks that are easy to achieve, to give the participants a sense of confidence and achievement. If tasks need to be made very difficult to complete successfully, participants need to be prepared for this, if possible, and certainly, participants should be appropriately debriefed (see

[11] Disability Rights Commission (2004). The Web: access and inclusion for disabled people. London: TSO.

Steps 5 and 6). The overall set of tasks should not be too long or tiring for any participant to complete (this can be determined in Step 5).

6.3.1.3 Step 3: Define what needs to be measured

Great care needs to be taken to ensure that what is measured in an evaluation meets the needs of purposes of the evaluation and that the results are really analysable (see comment in Step 5). The following kinds of measures can be considered (numerous classifications of such measures are possible, this classification seems helpful when practically planning the evaluation):

- Measures intrinsic to the task – e.g. time taken to complete the task; whether it is successful or not (if that makes sense in terms of the type of task); errors/confusions made when undertaking the task, problems encountered when undertaking the task.
- Participants' attitudes to the task – participants' comments about the task (made either concurrently, while doing the task; or retrospectively, after completing the task); participants' ratings of aspects of the task, e.g. ease of undertaking the task.
- Participant's attitudes to the whole system or website – after completing a series of tasks, participants can be asked more global questions, either open-ended or rating scales, to measure their attitudes to the whole system or website and their experience of using it.

In addition to planning what will be measured, one needs to plan how the measurements will be made – will the sessions be videoed, audio-taped, will an evaluator make notes, etc.

6.3.1.4 Step 4: Develop the overall protocol for the evaluation sessions

Once the tasks and the scenarios are planned, a protocol for the evaluation sessions with the participants needs to be planned. This should include:

- Material to be provided before the evaluation to the participants, to explain the evaluation to them.
- Consent and confidentiality forms.
- A script for introducing the scenario, the tasks, the questions, any other components of the evaluation session.

- Appropriate debriefing script and materials.

6.3.1.5 Step 5: Pilot the protocol

Even the most carefully planned evaluation will need a pilot to ensure that all contingencies have been considered and that the plan is feasible. Pilots can sometimes be conducted with several people who are not members of the target participant population, but wherever possible, the participants in the pilot should be members of the target participant population. The pilot should also be conducted in circumstances as close to the real circumstances of the evaluation, to ensure that all aspects of the evaluation are scrutinized.

A pilot can be used to investigate:

- Does the scenario make sense to the participants?
- Are the tasks achievable in a reasonable amount of time?
- Are the data collected really going to answer the issues of interest?
- Are the data collected going to be analysable?

6.3.1.6 Step 6: Recruit participants

An appropriate range of participants should be recruited. As with previous researchers, we have found that for formative evaluations, three to five participants from each user group of interest to an evaluation are sufficient to identify major issues. If normative data are required, larger numbers will need to be recruited. If particular differences between conditions or systems/websites are required, power calculations can be made to work out how many participants are needed.

We believe that participants should always be reimbursed appropriately for their time, as well as having travel and subsistence expenses paid.

6.3.1.7 Step 7: Run sessions

Sessions should be run with full and appropriate consideration to the participants. The length of time working with a computer should never exceed 50 minutes, and for participants who tire easily, should be less.

The session should not start unless the participant is completely happy and comfortable with the situation. It should always be emphasised that the participant has the right to withdraw from the evaluation at any stage (without financial loss) and that it is the system or website that is being

evaluated, not them. If the session is being video or audio-recorded, permission must be sought from the participant to do so, and the precise purposes that any video/audio material would be put to explained to them.

When the evaluation work is completed according to the protocol, sufficient time should be taken to debrief the participant and explain all aspects of the evaluation to them that they are interested in. The evaluation should be an interesting and potentially informative session for the participant.

6.3.1.8 Step 8: Analyse data

Data should be analysed in ways that protect the anonymity of the participants. Direct quotes from participants should not reveal personal identities. A pilot should have established that the data will provide answers to the evaluation issues when analysed.

Some data, such as classification of user problems, may require inter-coder reliability analysis.

7 Framework for evaluations of usability of existing accessibility E&R tools

7.1 Introduction

Coordinated by City University, we are conducting work on the functionality, usability and reliability of a range of E&R tools for Web accessibility. In this section, we will provide a brief overview on the methodology we are using for the usability component of the study.

Thus far, we have assessed the usability of three tools (Bobby 5.0 Desktop version,[12] Hermish,[13] and Wave 3.0[14]). There has not been a particular logic to the selection of these tools, simply that we had access to them. We plan to extend our study to as many tools as possible, and we may require assistance in getting access to some of the more expensive tools for long enough to conduct the evaluation.

7.2 The Methodology

7.2.1 Rationale for the methodology

The methodology we have chosen for this study is a heuristic evaluation (HE), a technique in which usability experts assess an interface (it can also be conducted by non-experts, but this is somewhat controversial and we used experts). We share the reservations of many HCI researchers about HE (Cockton et al, 2003), but we decided that was not feasible in this instance to proceed directly to user-based evaluations for the following reasons:

- problems of user recruitment:
  - for realistic user testing, one would need professional Web developers; these people tend to be well paid and very busy – finding enough people who could give up enough time to perform a serious evaluation would be extremely difficult
  - one factor on which we want to assess the E&R tools is how well they support Web developers who have little knowledge of

[12] http://bobby.watchfire.com/bobby/html/en/index.jsp
[13] http://www.hermish.com/
[14] http://wave.webaim.org/

Web accessibility issues, so for each tool we would have to recruit a new set of "naive" Web developers

- training overhead: some of the tools are quite complex and cannot be considered "walk up and use" systems; asking participants to spend time learning how to use a particular tool will increase the time required to conduct the evaluation, the commitment required by the participants (will probably require work outside the evaluation lab) and hence the difficulty of recruiting participants.

Therefore, to overcome these problems an HE method was shown. Classic HE is described in some detail by Nielsen (1993). Basically, a number of evaluators work through a system, looking for all the usability problems they think might exist. They may be given a scenario of use to work through and a profile of the type of user(s), or they may not. The theory is that any one evaluator will not spot all the usability problems, but if you pool the problems from 5—7 evaluators, you should pick up most of the problems. The evaluators should initially work alone, finding problems. They then have a group session where they discuss all the problems found, come up with a unified list and a set of severity ratings for the problems (see Table 1 and Table 2, below) which helps guide further development work.

The severity of a usability problem is a combination of three factors	
	The **frequency** with which the problem occurs: Is it common or rare?
	The **impact** of the problem if it occurs: Will it be easy or difficult for the users to overcome?
	The **persistence** of the problem: Is it a one-time problem that users can overcome once they know about it or will users repeatedly be bothered by the problem?

Table 1 Factors affecting a usability problem.

Finally, of course, one needs to assess the **market impact** of the problem since certain usability problems can have a devastating effect on the popularity of a product, even if they are "objectively" quite easy to overcome. Even though severity has several components, it is common to combine all aspects of severity in a single severity rating as an overall assessment of each usability problem in order to facilitate prioritising and decision-making.

The following 0 to 4 rating scale can be used to rate the severity of usability problems:	
0	I don't agree that this is a usability problem at all

1	Cosmetic problem only: need not be fixed unless extra time is available on project
2	Minor usability problem: fixing this should be given low priority
3	Major usability problem: important to fix, so should be given high priority
4	Usability catastrophe: imperative to fix this before product can be released

Table 2 Severity ratings in heuristic evaluation.

We realised that the problem with the classic HE method for us was still the training overhead in understanding some of the tools. So we have developed a variation of the method, which seems to have worked quite well. We call this the "one step" HE.

7.2.2 One step HE

All the evaluators sit around the system together (we found it best if the Web page was projected on a screen, so we could all see easily). One person is the nominated "driver" of the system – the person who understands it (or understands it the most). They demonstrate how to work the system and answer questions about its use.

Anyone is allowed to nominate usability problems, which are then discussed until everyone is clear about the problem and it is given a name and number. Then each evaluator states privately on their rating sheet whether they actually think it is a usability problem and give it a rating.

We hope with this method we have overcome the training overhead problem without compromising the elicitation of usability problems.

We have conducted two evaluations with the one step method and one with the more usual two step method. As the evaluations were on different systems, it won't be possibly to directly compare the methods themselves (if the one step method produces more usability problems it may simply be because that system had more problems), but it should give some feel for the methods.

8 Conclusions

The deliverable has presented an overall approach for a framework to monitor different aspects of the project. During BenToWeb's lifetime, this framework will be refined and tailored to the different elements under scrutiny. The whole endeavour is a collaborative and iterative effort among all WPs involved, and we expect to update this document periodically with further information arising from our experiences.

9 References

Caldwell B, Chisholm W, Vanderheiden G, White J (2004). Web Content Accessibility Guidelines 2.0, W3C Working Draft 19 November 2004. World Wide Web Consortium (W3C). Available at: http://www.w3.org/TR/WCAG20/

Chisholm W, Vanderheiden G, Jacobs I (eds) (1999). Web Content Accessibility Guidelines 1.0, W3C Recommendation 5-May-1999. World Wide Web Consortium (W3C). Available at: http://www.w3.org/TR/WCAG10/

Cockton G, Lavery D, Woolrych A (2003). Inspection-based evaluations. In: Jacko J, Sears A (eds), The Human Computer Interaction Handbook. Mahwah, NJ: Lawrence Erlbaum Associates.

Nielsen J (1993). Usability engineering. New York: Academic Press. See also: http://www.useit.com/papers/heuristic/heuristic_evaluation.html