

Developing Usability Tools and Techniques for Designing and Testing Web Sites

Jean Scholtz and Sharon Laskowski

National Institute of Standards and Technology (NIST)

Bldg, 225, A216

Gaithersburg, MD 20899

1-301-975-2520 1-301-975-4535

Fax: 1-301-975-5287

Jean.Scholtz@nist.gov Sharon.Laskowski@nist.gov

Laura Downey¹

Vignette Corporation

3410 Far West Blvd., #300

Austin, TX 78731

1-512-425-5495

ldowney@vignette.com

Keywords:

Usability testing, web development, web site design, remote usability testing, web site evaluation.

Introduction

Web sites are developed by a wide variety of companies from small one or two person operations to large corporations with entire teams and departments devoted to web development work. Many of the smaller companies have no usability professionals to help with the design of their sites; many of the companies don't even realize they should have usability professionals assess their sites. Often, budget constraints prohibit hiring a usability professional. Furthermore, the time line for development may not allow for usability testing and iterative design.

Traditional usability methods are expensive, time consuming, and many require professional usability engineers. The environment in which web sites are designed cannot easily support these constraints. Our approach assumes that evaluation of web sites must be rapid, remote, and as automated as possible. In this paper, we'll discuss three software tools and two techniques that we have developed to facilitate quick evaluation. These tools and techniques can be found at <http://zing.ncsl.nist.gov/~webmet/>. We'll also describe supporting case studies and our plans for future tools and techniques.

¹ This work was completed while this author was a NIST employee.

Approach

We are currently developing software tools and techniques for evaluating the usability of web sites. Our approach includes two types of tools and techniques. The first we'll call Usability Awareness tools. These are tools for use by web site designers and developers who may not be familiar with usability issues. These tools should educate web developers on usability issues in general and the specific instances of usability issues in their sites. The second set of tools and techniques we call Web Usability tools; tools developed for use by usability professionals. In developing this set of tools we are concerned with increasing the effectiveness of usability professionals working on web sites and applications. To achieve this we are focusing on developing tools and techniques to speed evaluation (rapid), tools that reach a wider audience for usability testing (remote), and tools that have built-in analyses features (automated). Our approach is to develop these tools and techniques based on case studies we have done on various web sites and applications. We used the information gained in these case studies to design the first version of the tool. Once the tool is developed, we will use information from applying the tool to different web sites and applications to identify new functionality as well as identify limitations of the tools. We have also conducted case studies to produce techniques that are rapid, remote, and automated. Some of these techniques may eventually become tools but for some, it is sufficient to describe the technique so that usability professionals may apply this to their own sites.

Usability Awareness Tools

In carrying out case studies, we try to identify tools that could be developed to help in the design or testing of web sites. In this section, we describe an automated tool we developed for evaluating usability aspects of a web site

A Software Tool: WebSAT (Static Analyzer Tool)

There are currently several tools available on the web to check the syntax of html. Tools also exist that identify and check accessibility issues with web sites. The first tool we developed carries this concept one step further by providing an easy way to check potential usability problems with a web site. WebSAT (Web Static Analyzer Tool) is one of the tools in the NIST WebMetrics suite. To develop WebSAT, we looked at many of the design guidelines for the web and some of the case studies that have been documented (Detweiler, Spool). Some of these guidelines can be checked by looking at the static html code. For example, many accessibility issues can be addressed by looking at such things as the number of graphics that contain ALT tags that can be read by screen readers. We can check that there is at least one link leading from a page as novice users may not understand how to use the browser "back" button. We check that the average number of words used in describing links is adequate but not overly wordy so that users can easily find and select a promising link. For the readability of a page, we can check how much scrolling text or marquee style text is used. In addition to performing such usability checks, WebSAT contains a table to explain what is being checked and why this might constitute a usability problem.

Future plans include the ability to look at an entire site at one time, as opposed to the page-at-a-time view currently supported. Interactions between pages account for many

usability problems, some of which we may be able to predict from static text. For example, a check could be made to compare the similarity of the link text with the heading on the page it links to. We also plan to include an interactive section, where users may ask to have certain features checked, such as the inclusion of a particular graphic. This would be a way to check adherence to some corporate look and feel guidelines.

Web Usability Tools and Techniques

The NIST WebMetrics tool suite contains two tools to help usability professionals, WebCAT and WebVIP. WebCAT is a category analysis tool that allows the usability engineer to design and distribute a variation of card sorting (Nielsen), over the internet. WebVIP, a visual instrumenter program, provides the usability engineer with a rapid way of automatically instrumenting the links of a web site (identifiers and timestamps) in order to prepare the web site for local or remote usability testing (Hartson et. al.) and subsequent recording of the user paths taken during testing. Both WebCAT and WebVIP are targeted towards usability engineers as the construction of the tests and interpretation of the results need professional expertise. These tools facilitate data collection and analysis and support our goal of rapid, remote and automated testing and evaluation methods. We have also identified two techniques that we found useful in designing a web site and in usability testing a web site. We describe these two techniques as well. For each tool and technique described below, we first describe case studies we carried out to determine the feasibility of the tool or technique and to arrive at the initial requirements.

A Software Tool: WebCAT (Category Analysis Tool)

Case Study One: NIST Virtual Library (NVL)

We carried out two case studies to determine the feasibility of doing variations of card sorting remotely. The first case study was conducted with the NVL.

The NVL staff was considering a redesign of the web interface and was very interested in obtaining data that would help them focus on problem areas. We devised a usability test to identify problems with the existing design. One of the parts of our usability test was a matching exercise to test existing categorization. For example, the categories used on the NVL page are: Subject Guides, Databases, E-Journals, NIST Publications, Online Catalog, Web Resources, Hints and Help, NIST Resources, and Visiting. We recruited five subjects from different scientific disciplines who worked at the NIST site in Gaithersburg, MD. We wanted a baseline for comparison so we used two expert users from NIST, a reference librarian and the web site designer.

Our matching task was a variation of a traditional card sorting task. In a card sorting task, users supply names for actions and objects in the interface, group them, and then label the groups. As we were not starting from scratch in our design, we used a matching task. Our goal was not to determine new category labels but first, to determine if any categories were troublesome.

In the matching task users were asked to assign 29 items to one of 10 categories, nine categories from the NVL home page plus a “none” category. We scored users’ responses according to:

- The number of items assigned to an incorrect category
- The number of times a category had an incorrect item assigned to it.
- The number of users who assigned incorrect items for each category.

Our baseline users placed two of the 29 items in the wrong category. Our non-expert subjects placed an average of 13 items in wrong categories. This clearly indicated that the classification scheme used in the site was unclear to the users. Of the nine categories, all subjects had problems with three: databases, hints and help, and NIST resources.

It is important to note that in this study, the category analysis was not done remotely. It was done on paper with the experimenter and subject in the same room. However, we limited the interaction in this part of the evaluation as we wanted to "simulate" remote evaluation capabilities.

Case Study Two: Information Technology Web Site

Our second case study was the redesign of the Information Technology Laboratory (ITL) web site. ITL is one of seven laboratories at NIST. Again, we did a category matching exercise but this time the categories and the items were those that were under consideration for the redesign. In this case study we did perform the evaluation remotely. We e-mailed the category matching exercise to fourteen participants. Working with the web site designer, we set a baseline of 75% identification of items as our goal. The results were actually better than the goal we established. On the average users identified 87% of the items correctly, with only one user performing below the established criteria. In this case study, we also asked the participants to describe the categories we identified before doing the matching exercise. This helped us to get qualitative information about the meaning of the different categories.

The Tool: WebCAT

Based on these two case studies, we designed WebCAT. A usability engineer can use this tool to design a category matching exercise and then distribute it remotely to users. The results are automatically compiled as each user completes the exercise so the usability engineer can quickly assess any problems with the categories. The usability engineer uses WebCAT and specifies the categories and items or subcategories he/she is going to use. WebCAT produces a GUI interface for the test where the user drags items or subcategories to labeled category boxes to complete the exercise. The usability engineer uses this same method to produce the baseline or comparison case. After the baseline is completed, the usability engineer can send out the URL to usability participants.

Currently, WebCAT is undergoing beta testing and will be publicly released by the time of this conference. Future plans for WebCAT include an improved analysis program.

A Software Tool: WebVIP (Visual Instrumenter Program)

Case Study: NIST Virtual Library

WebVIP is a program that allows usability engineers to visually instrument an existing site so that users' paths can be tracked as they perform tasks specified by the usability engineer. This tool was developed following our case study of the NIST Virtual Library (NVL).

In the preceding section, we described the matching task we gave participants. For the task performance part of the case study, we concentrated on tasks that required users to locate information. Our goal was to see if we could collect a bare minimum of data and still identify usability problems. For the 10 representative tasks users were asked to do, we collected:

- Whether users found the answer (yes/no)
- The time it took
- Users' perceived difficulty
- Users' perception of the time for completing the task

Again, we used our two experts as a benchmark for comparison. Recall that we did not conduct this case study remotely but did limit the interaction between the experimenter and the participant to simulate "remoteness."

Each expert user was able to do 9 of the 10 tasks. However, each expert user missed a different task. Of the non-expert participants, three users successfully completed six tasks and the other two users successfully completed seven tasks.

Each expert user took just over eight minutes on average to complete the ten tasks. The non-expert users needed on average over 31 minutes to complete the same tasks. Looking at individual tasks, we found that all the non-expert users commonly missed one task. However, they did not rate this task as the most difficult. This is probably because many of them thought they had located the answer. This indicated to us that we needed to collect the answers to information seeking tasks to ensure that users were successful.

Users rated the difficulty and time factors for the tasks quite high given the success and time they needed to complete these tasks. Experts rated the difficulty of the tasks as 5.7 on a 7 point scale; where 1 was very difficult and 7 was very easy. The non-experts' average difficulty rating was 4.8. Average ratings of the time it took to accomplish the task were 5.8 for the experts compared with 4.8 for the non-experts. Again, these ratings were on a 7 point scale, with 1 being too long. The user ratings of task difficulty and time were very closely correlated. Thus it seems that a perceived difficulty rating for the task alone is sufficient.

Looking at the paths that users took to locate information gave us quantitative data about different strategies that users took. We were also able to identify areas that were misleading to users.

The Tool: WebVIP

Based on this case study, we developed WebVIP to use in tracking user paths for specified tasks (both location and time). The usability engineer needs supply a copy of the actual web site. This can be done by using one of several webcrawler programs. Then the copied site is instrumented; that is, code is added to the underlying html so that the links followed during usability testing will be recorded. Recording a link means that an associated identifier and time stamp will be recorded in an ASCII file each time a test subject selects a link during usability testing. WebVIP also lets the usability engineer designate links as internal (links to other pages within the site) or external (links to pages outside the site). This is useful in determining if users are spending more time within the test site or going outside the test site seeking answers to the tasks. It should be noted that WebVIP only records path information for the instrumented site; it does not record anything when users leave the instrumented site. Finally, the usability engineer can also record and associate comments with links and/or the site itself. These comments will be recorded in the text file along with the link name and time stamp when the link is traversed or in the case of a site comment it will merely be recorded at the top of the file. A small "start/stop" graphic is also added to each page in the web site so that the user will have a method for indicating the start/stop time of each task during usability testing. From the text file containing the data about traversed links, the usability engineer can identify time on task, time on particular pages, use of the back button, links followed from pages, and the user path taken for the specified task.

As with WebSAT and WebCAT, WebVIP will be publicly released by the time of this conference. Future plans for WebVIP include an analysis program with visualizations of the user results as well as enhancements to capture qualitative data, e.g., a prompt box specified by the usability engineer that will appear when a particular link is followed during usability testing to record the user's current thinking at that decision point. We also plan to allow usability engineers to incorporate their test directions into the instrumented site.

A Technique: Using Beta Testing to Identify Usability Problems in Web Applications**Case Study: NIST Technicalendar**

A printed calendar (the Technicalendar) is published every week at NIST. It contains notices about meetings and talks to be held at NIST, notices of talks given by NIST employees at other locations, as well as meetings elsewhere that might be of interest to NIST scientists. The calendar is distributed to NIST personnel in hardcopy. It is also viewable on the web and e-mailed to others outside of the agency.

Previously, articles for inclusion in the Technicalendar were faxed, phoned in, or e-mailed to a staff person. This person spent at least one day per week collecting any missing information for items submitted and formatting them correctly. To streamline this activity, an on-line wizard was developed so that submissions could be made via the web. It was hoped that this would considerably reduce the time spent in publishing the Technicalendar and make the submission process easier for both professional and administration personnel.

We decided that a beta test might be an appropriate way to collect usability information. We provided three ways to identify usability problems. We constructed an evaluation form for users to fill out after they had used the Technicalendar Wizard. We provided nine rating questions about usability of the form including navigation between fields, navigation between pages, optional versus required fields, and terminology. In addition we included two open-ended questions for users to tell us about any special usability problems they encountered. We also gave users the option of submitting a "test" item or a real item. The test submissions and the real submissions were available to us to use in identifying usability problems.

Case Study Results

The entire test period lasted 10 weeks. During that time we received 83 submissions; 59 of these were actual submissions and 24 were "test" submissions. We received 28 completed questionnaires. Due to the time lines for design, we assessed the site after one month of testing. During the first month, we received 24 submissions, 16 of these were actual submissions and 8 were "test" submissions. We received 13 questionnaires.

What problems were identified in what ways? Table 1 lists the three methods and the types of problems that were identified using each method. The open-ended comment section provided the most information about usability problems. This section was particularly helpful in providing information about special cases (panels with six participants, providing special formatting for special cases, etc.).

<i>Identification Method</i>	<i>Type of problem</i>
Calendar submission	Text field formatting
Low ratings	Determining optional fields
User comments	Access to help
	Relationship between fields
	Terminology
	Layout
	Missing defaults

Table 1: Ways in which usability problems were identified during beta testing

We worked with the WebMaster to correct the problems identified and a second version of the on-line submission wizard was installed on the web site. We continued the beta

test to see if the changes had really corrected the problems. During the next six weeks, we received 59 more submissions, 43 of which were actual submissions and 16 were "test" submissions. We received 16 questionnaires as well. We did not uncover any new problems in the second round of testing. And in this case, we were actually able to use the ratings in the usability questionnaire to verify that our redesigns resulted in improvements. A side benefit of the beta testing and formal evaluation was that exactly one complaint was received when users were required to use the wizard for all Technicalendar submissions.

Technique Recommendations

Our recommendations for using beta testing to identify usability problems include:

- Beta testing is useful for small, focused web applications.
- Provide a short rating scale for participants that can be used to validate that usability problems have been fixed or improved.
- Provide a way for users to make open-ended comments.
- Provide a "test" option so users can use the software even if they do not currently have "real" data or information to provide.
- Collect data or information from the test cases as well as the real information.

We plan to apply this technique to different types of web sites to see how useful it is in information seeking sites. We anticipate that this may be useful in sites where the tasks are limited but less useful in large, multipurpose web sites as usability problems occur in context. It is possible to deduce users' tasks in small, limited use web sites, but not in large, multipurpose sites.

A Technique: A Virtual Participatory Design Meeting

Case Study: Identifying Requirements for the NIST Virtual Library

Our second technique was developed as a result of the redesign of a major NIST on-line resource – the NIST Virtual Library (NVL). The NVL is a scientific library accessible to the public from the NIST web site. While some of the databases are restricted to NIST personnel, most of the library resources are open to the general public.

One of the interesting issues with a library site is that there are two very different categories of people who need to be considered when redesigning such a site. While we are, of course, interested in the library user and how well the site meets their needs, we also need to consider the impact of redesign on the library staff. Much behind the scenes work is still needed to make a virtual library "virtual." We wanted to ensure that we considered first of all, input from library staff for the redesign. The library staff already has a tremendous amount of work and due to the various hours they work, scheduling meetings is difficult. But it was important to make sure that we gathered as much input from them as possible. We also felt that people are much more likely to think of issues affecting the design one issue at a time- and these ideas arise because of something that is

happening at work that moment. How could we make sure that we captured that information in a timely way?

We decided to collect scenarios from the library workers via the intranet as all of the staff has easy access to this during their work. We scheduled an initial meeting with the staff and explained what we wanted to do. We then had a period of several weeks in which the staff contributed their scenarios. We supplied a template for the scenarios. We allowed the staff to comment on others scenarios anonymously and also made it easy to see which scenarios others had commented on. We supplied several example scenarios to help everyone get started. E-mail notification was used to alert staff when a new scenario was posted.

For each scenario, the submitter was asked to include a description, identify the benefits (speed, accuracy, not currently possible, etc.), and who would benefit (library reference staff, end users, other library staff, etc.). We asked the submitter how frequently this scenario would happen and its importance on a 7-point scale. We also asked for any negative aspects to this scenario.

Case Study Results

We received 28 scenarios. Of these, 18 also included comments from one or more participants. After the collection period was over, we classified the scenarios into basic categories, with scenarios being allowed to be in more than one category. We are using these categories and scenarios to construct initial requirements for the revised NVL and we will use these scenarios in the design of our usability evaluations. We also plan to collect scenarios from end users as we redesign the end user portion of the virtual library.

Technique Recommendations

- Use this technique with users who feel comfortable using electronic communication and have easy access to it.
- Provide examples so users understand how to fill in the blanks.
- Use e-mail to alert users to the arrival of new scenarios. This reminds them to enter their own information.
- Encourage several users to be quite active in submitting information and commenting early in the process as this encourages others.
- Make sure that users know how this information will be used in the redesign.
- Pilot test the scenario template form with several participants to make sure it is easily used and understood.

Future Plans

Our WebMetrics site (<http://zing.ncsl.nist.gov/~webmet/>) contains descriptions of the tools, downloadable code, and examples of the use of the tools. We are in the process of adding descriptions of the techniques to the site.

Now that we have produced these tools and techniques, we plan to use them on many different types of sites to determine the type of sites for which the tools and techniques are most useful. Validation of these tools is an important aspect of our work, but one that

we alone cannot carry out. We hope to obtain feedback from users of our tools about their experiences. We are interested to hear about the types of sites where these tools and techniques are used and if the tools and techniques were helpful in producing more usable web sites.

We have several other software tools that we are currently working on as well as providing improved functionality for the existing tools. We are currently developing visualizations for server log data to provide comparison of user paths. Such a tool will allow usability engineers to see the effects of redesigned pages on user paths, compare cultural differences in use, and view the change in usage patterns over time.

We are also carrying out a kiosk-based usability evaluation. This involves the design of an engineering statistics handbook. We are interested in seeing the type of usability information we can obtain in very short evaluation sessions.

We will continue our approach of doing case studies, developing tools or techniques based on the results of these case studies, and then applying these tools and techniques to different types of web sites. Our final goal is to produce a systematic methodology, including tools and techniques, to facilitate the production of usable web sites and applications.

Acknowledgements

Our thanks to Charles Sheppard, Paul Hsaio, Dawn Tice, Joe Konczal, and John Cugini for their help in conducting the case studies and in product development and testing.

References

Detweiler, M. and Omanson, R., Ameritech Web Page User Interface Standards and Design Guidelines,

http://www.ameritech.com/corporate/testtown/library/standard/web_guidelines/index.html

Hartson, H. Rex, Castillo, Jose, Kelso, John, and Neale, Wayne. 1996, Remote Evaluation: The Network as an Extension of the Usability Laboratory, in *CHI'96 Conference Proceedings*, April 13-18, Vancouver. 228-235.

Nielsen, J., 1993, *Usability Engineering*, Academic Press, Boston.

Spool, J., Scanlon, R., Schroeder, W., Snyder, C. and DeAngelo, T., 1997. *Web Site Usability: A Designer's Guide*, User Interface Engineering.