Evaluating Web Site Structure
A Set of Techniques

K. Frederickson-Mele, Michael D. Levi, and Frederick G. Conrad
U.S. Department of Labor, Bureau of Labor Statistics
Washington, DC

Introduction

As the novelty of World Wide Web site development subsides, old lessons are resurfacing. Once again the importance of user centered design is becoming increasingly evident. Given the sheer volume of Web sites and Web users, the opportunities for wasted effort and time are overwhelming. Thus, creating sites that are compatible with the way people organize, remember, and use information is essential.

Web site design can be broken into two main components: page design and site structure. Page design is roughly comparable to screen design in other interactive environments, and is concerned with effective and consistent arrangement of screen objects; appropriate use of language; proper utilization of color, typefaces, and graphics; etc. Site structure is concerned with the proper organization of content to facilitate efficient and predictable navigation. Inter-page and intra-page movement is the locus of interactivity in most Web sites (leaving aside forms or scripted applets); thus site structuring largely takes the place of dialog design in other types of interactive software.

Many rather good Web style guides are appearing, which can help tremendously in page design. Site structure, however, is much more difficult to generalize, as dialog and navigation are so tightly linked to the details of any given task. Validating a proposed page design is important, validating a proposed site structure is essential.

Background for the Case Study

The Bureau of Labor Statistics (BLS) released an Internet Web site in September, 1995. A primary reason it was successful was because its usabilty was evaluated repeatedly, beginning with the prototype site (Levi & Conrad, 1996). Encouraged by this experience, the BLS decided to use the new Web technology to develop an improved procedure for distributing internal information to its employees.

A small intranet design team was established in August, 1996. The team was given two weeks to design an approach for developing an intranet, and present recommendations to upper management. Their proposal recommended that a prototype be developed, and be subjected to usability testing.

The prototype was completed in October. Management approached the usability test team, and requested that usability testing be conducted. The test team, which consisted of the authors, was given two and one half weeks to design the evaluation, conduct the tests, evaluate the results, and present the findings. We met with the prototype's management team to discuss their requirements, and identified the overall organizational issues (rather than the content of the individual pages) as being the most critical.

Given this focus on the prototype's high level structure, we decided to use a Card Sort Exercise, an Icon Mix-and-Match test, and a Category Membership Expectations test. The Card Sort exercise was designed to determine what mental hierarchy users construct when given a set of anticipated leaf pages from the intranet site. The Icon Mix-and-Match test was designed to find associations and/or interference between button pictures and the associated button text. The Category Membership Expectations test was designed to elicit users' understanding of a set of categories and their associated labels. The tests were conducted in one afternoon, in two separate sessions (first the Card Sort, then the Icon Mix-and-Match and the Category Membership Expectations test).

Seventeen test subjects were identified and recruited. None of the participants were involved in designing or implementing the prototype. They were distributed over as many offices as possible. All were expected to have worked in the Bureau long enough to have a reasonably firm grasp of the organizational structure, and the type of work performed there. Some Web use experience was preferred. Since the objective of this series of tests focused on the prototype's overall design, the tests did not address individual page design.

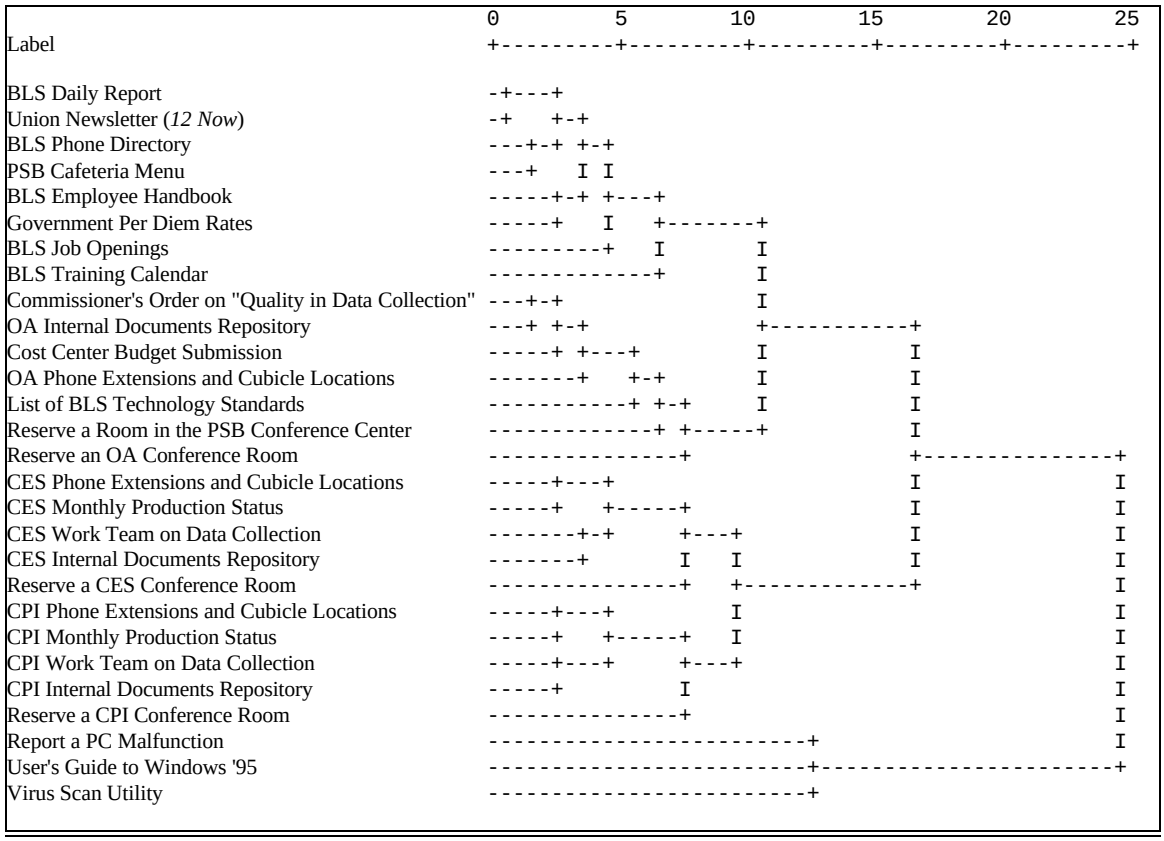
Conducting the Tests

The Card Sort

Each participant was given one set of randomly ordered colored index cards that contained the individual items (anticipated leaf pages from the intranet site), rubber bands, and blank white index cards. They were asked to arrange the items into logical groupings and place a rubber band around each group. If the banded groups could be further aggregated, they were asked to band those, place a blank white index card on top, and label each grouping with a title that best described its content.

Sample items (anticipated leaf pages)
BLS Phone Directory
Reserve a Consumer Price Index (CPI) Conference Room
Consumer Price Index (CPI) Monthly Production Status
Consumer Price Index (CPI) Work Team on Data Collection
Reserve a Current Employment Survey (CES) Conference Room
Current Employment Survey (CES) Monthly Production Status
Current Employment Survey (CES) Work Team on Data Collection

We performed a hierarchical cluster analysis on the data. The resultant dendrogram aggregated the users' sorting decisions, providing a picture of how the users mentally organized the site's (likely) information. This was not the final word on an optimal site structure. By comparing the results to the proposed site structure, it was possible to see how well it reflected users' mental organization of the same information. In this card sort, the participants could have grouped leaf nodes by function. Some respondents, for example, did group all conference room reservations together, but the majority grouped leaf nodes by the office or program. This largely confirmed the design assumptions of the prototype.



Card Sort Dendrogram

In addition to looking at the composite results, we attempted to determine how or why individual users may have sorted the cards as they did. One pattern in the individual users' piles was they clearly distinguished between agency wide functions and the functions of their particular office or division. We return to this in the section "What Happened as a Result."

The expenses associated with card sorting were the time it took to make the cards (one day), conduct the test (an afternoon), and compile the results (two days). Conducting this test would have been less useful if the test participants had been new employees, and consequently less familiar with the agency's organization. Card sorting can be used anytime in the development process when information needs to be "chunked". However, when used early in the design phase, it helps produce a well organized site from the start; this avoids wasted effort designing individual pages that may become obsolete if the structure is changed later.

The Icon-Mix-and-Match

Research in human-computer interaction has found that the benefit of an icon/label pair is that the two different formats reinforce one another [1]and users can focus on either the picture or the text, whichever is more efficient. To make this possible, people must agree that a particular icon brings to mind the same concept that a particular string of text brings to mind, and that the icon does not bring to mind other concepts. We tested this for the intranet site with an "Icon-Mix-and-Match" test.

In this test, participants selected an icon they felt best represented a category. "What's New," for example, might be matched to a tiny picture of a newspaper by a participant choosing from a matrix of pictographic representations. If several other participants made the same choice and did not pick the newspaper to represent something else, then there's a good chance that the wider user population will make that association as well. The test team looked at how participants matched a textual category label to an icon, as well as any possible interference existing between icons and category labels. We established a threshold of 70% agreement for an icon label pair to be successful.

Participants were asked to match 16 icons with six categories. Participants were given a table with the icons as column headers and the categories as row stubs. They were instructed to select the best icon that represented the corresponding category, and place an X in the cell. If more than one icon corresponded to that category, or none did, participants could place an X in multiple cells, or in none. The icons and the categories were placed in random order to minimize bias.

Category	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)	(n)	(o)	(p)
REFERENCE docs & materials			10		1		1		4	1					2	
TOOLS applications		3			1		1	10				2				2
ORG VIEW internal structure		1		7							10				2	
SERVICES support functions					7							1	2			
NEW what's the latest	10									10						
MAP site menu						7								10		

Icon Mix-and Match matrix

The strongest match for any category was 100%, the weakest was 10%. The criterion for selection was a category with a match equal to or greater than 70%. The test team also looked for any icons that matched multiple categories, where this was defined as an icon being assigned to two different categories 40% or more of the time. This never occurred.

Minimal resources were required to conduct the test. We designed the spreadsheet in less than one day, and it took approximately 20 minutes for the participants to complete, and about an hour to tally the results. Since this

[1] Kent L. Norman, *The Psychology of Menu Selection* (Ablex Publishing, 1991)

particular test had icons and possible associated categories, it was not sensitive in determining whether users might have provided other textual category labels for the icons.

Category Membership Expectation

A Category Membership Expectation test is designed to elicit users' understanding of a set of categories and their associated labels, and thus determine how or if an organizational scheme is usable. We looked for thematic agreement among the participants as to what they believed would appear in a certain category, and whether or not it conformed to the intranet designers' view. Responses were tallied and consolidated.

Participants were given a form which listed the six prototype categories. They were asked to list the kind of material they would expect to find in those categories when they were on the BLS home page, and when they were on their organization's home page. To preserve context, the categories were listed in the same order as they appeared in the prototype site.

The categories were: (the phrases in parentheses explained the category to the reader)

1.New (what's the latest)
2.Reference (docs & materials)
3.Tools (applications)
4.Services (support functions)
5.Org View (internal structure)
6.Map (site menu)

Two categories clearly worked as intended: 'New' and 'Reference'. Most people had clear ideas of what topics they would expect to see in these two categories, (all participants expected to see updated or new pages under “New” and there was agreement among the participants at both the BLS and organizational levels with respect to reference.) However, three categories did not work: 'Tools', 'Services', and 'Map'. There was no agreement as to what should be placed under any of these categories, with only three or four participants listing items that conformed to the intranet design team’s definitions. The remaining category, 'Org View', was a partial success: users expected an organization chart, but not a clickable chart that could be used for navigating to sub-sites.

Minimal resources were required for this test. We designed the form in less than one day, and it took approximately 30 minutes for the participants to complete, and about three hours to tally the results. This test method should be used early in the development process.

What Happened As a Result

One unexpected result -- due in large part to organizational politics, but brought into perspective by the usability tests -- had to do with control of the intranet as a whole. The initial expectation of the prototype development team was that the entire site would be internally consistent, with a high-level structure mirrored in program-specific subsite structures. A single organizational model was to be implemented by each office. Development and maintenance would be centrally controlled. The prototype organized material by the originating office.

The card sort, in particular, showed us that users did not quite follow this mental model. Instead they distinguished between BLS-level information (such as personnel policies or technology standards, regardless of the source) on the one hand, and program-specific information (such as Consumer Price Index meetings) on the other. It seemed likely that the majority of users would access the BLS-level information, and their home office information, but would rarely if ever access another office's internal material.

We emphasized the importance of consistency between the high-level structure and any given office's organization, but also pointed out that cross office consistency was much less significant.

This made the intranet management team's political task much easier. No longer would offices be forced into consistency with one another; instead they merely needed to maintain commonality with the central site. Control and content creation was subsequently decentralized.

We also strongly recommended that further page level and sub-site level testing be carried out in the future. We noticed many pages within subsites that we considered sub-optimal. However, management has not requested a follow-up study.

Observations and Reflections

When we originally met with the management team, we used an interview script. This assisted us in deciding to focus on the prototype's high level structure. Given the two week time constraint, we used techniques that could be applied rapidly. This meant that we could not collect data from end users visiting the prototype site because that kind of study is relatively slow and labor intensive. All tests were conducted using paper materials which were constructed quickly. The tasks were quick for the participants to complete and were relatively quick to analyze. (The card sort, however, involved a laborious data entry process.)

Recruiting the participants was conducted by a source outside of the evaluation team, and as a result, there were some communication problems. For example, we were compelled to disqualify some of the participants because they were part of the intranet effort. Similarly, three participants were recruited for all of the tests even though we intended to use one group for the card sort, and another group for the other two tests. Consequently, we believe that we should have been very specific in our participant profile request. A final issue concerning participants is the breadth of the sample. While the skill mix was diverse with respect to web experience, we could have expanded the participant base to include users from additional offices.

After the tests were completed, we presented our findings to the management team. If we were to do this again, we would not use the dendrogram as an explanatory technique. Many individuals found it difficult to interpret. A final point is that our recommendations should have been more explicit. We may have expected readers of the final report to infer too much.

Conclusion

The *interrelationship* between the tests is what made them especially useful because each asked users to group essentially the same type of information, (categories), but through different instruments. The Card Sort asked participants to group predetermined items into hierarchical categories, while the Category Membership Expectation test asked users to place their own explicit items in the categories. The Icon Mix-and-Match asked participants to match icons with categories. The tests results provided valuable information about the proposed categories. The Card Sort exercise validated the prototype designers' fundamental approach: follow the BLS organizational structure. The Category Membership Expectation test showed that the specific instantiation chosen by the designers was flawed, at least for some of the categories. The Icon Mix-and-Match identified icons that were not interpreted as the designers expected they would be.

Our interpretation of the results lead to a high level split between information of general interest to most BLS employees, and information relevant to specific offices.

This series of tests did not address page design at all. The evaluators noticed many areas in different pages and sub-sites that we considered sub-optimal (gratuitous animation, for instance). Hence, we strongly recommended that page level and sub-site level usability testing be carried out in the future.