Foundations Project Usability Testing: Dublin Core Metadata and Controlled Vocabulary Study

Introduction

The Foundations Project is a State of Minnesota multi-agency collaborative project developed to facilitate the access to environmental and natural resources data and information from 13 Minnesota state agencies through the Project's **Bridges Search Interface**. Project leaders developed metadata cataloging guidelines and other searching aids designed to be intuitive and easy to use for both specialists and non-specialists. Project staff and agency participants trained by project staff have added Dublin Core metadata to electronic data and information resources including web pages, PDF documents, tabular data and geographic data. Advanced search and retrieval techniques that integrate access to this information across agency websites have also been designed. The Ultraseek search engine is employed.

Goals

One goal of the Project is to increase accuracy and relevancy of search results from the Ultraseek search engine. The goal of this Usability Study is to test keyword searches of web pages to determine the effect of controlled vocabulary in the Dublin Core subject element (dc.subject), comparing controlled terms from the chosen thesaurus with uncontrolled terminology or synonyms. The analyst looked at the change in rates of dead-end searches, excessive results, and irrelevant results with the addition of this metadata element.

Background

The crisis in searching the World Wide Web has been evident for some time. A November 1999 Internet conference, in their "What's Hot and New" presentation, looked at the need for the human element in getting control of the vast amount of searchable information on the Web. The *New York Times* "Circuits" section from June 29, 2000 headlined the need for human intervention as well: "As the Web sprawls out of control, search engines are overheating and programmers are trying something new: human beings." Librarians and indexers have been aware of this need for some time, and now many major search engines, such as Yahoo!, Northern Light, and Google, are employing information professionals to improve search results. The Foundations Project made a decision early on to provide embedded metadata, applying such standards as the Dublin Core metadata element set, controlled vocabulary from the Legislative Indexing Vocabulary, Minnesota edition (LIV-MN) thesaurus, ISO standards for such elements as date and language, and a uniform method for entering personal and corporate names. It is the belief of the Project leaders that consistency in application of the 15 Dublin Core elements would lead to more accurate and relevant search results for even the novice searcher.

During the course of the application of metadata to Foundations Project participants' web pages, an outside consultant has been testing and analyzing a group of keywords (uncontrolled vocabulary) and thesaurus

terms (controlled vocabulary). The test was conducted in three passes:

- Pass 1 looked at all terms as keywords, or uncontrolled terminology, as there was no metadata on the pages.
- Pass 2 looked at the same search terms (by now a mix of uncontrolled terminology and thesaurus terms) where metadata had been applied to about half of the appropriate documents.
- Pass 3 looked at the same search terms when the Project was very nearly complete, and metadata had been applied to nearly all appropriate documents.

Methodology

Testing involved creating a methodology, including spreadsheet to record results, to be used for simple searching – search terms were either single words or phrases, either from the chosen vocabulary or keywords. Comparisons of the three search passes and their results helped determine the success of the metadata project.

Information recorded includes:

- Searches that had no results ("dead-end searches").
- Searches with over fifty retrievals.
- Searches that resulted in an insufficient number of hits.
- Searches that found inappropriate sites.
- Search results using designated keywords.
- Search results using designated terms from the Legislative Indexing Vocabulary.

Also included in the original spreadsheet are:

- A "qualitative comparison" section, which reports relevance ranking for the first and last documents in the top fifty results, as judged by the Ultraseek search engine.
- A "format" or "pdf" section, which the analyst used to record any notes about format, including documents in Adobe Portable Document Format.
- A "time" section, which was later deleted as irrelevant.

As the project developed, improvements in the Ultraseek search engine, the infrastructure behind it (such as the use of spiders/robots), as well as increased numbers of agency Web sites with metadata, combined to result in more effective searching. This is evident in the analyst's comments on the changes from the first testing pass to the second and third. Improvements in Ultraseek included allowing the use of truncated search terms, more consistency in total returned hits and better help screens. Over time, response time improved as well, with shorter waits for downloads.

Conducting the Test

The analyst contracted for the purpose conducted all searches. Steps included:

- Simple search for each designated term.
- Quoted search for phrases.
- Capture top 50 results to file.
- Clean extraneous material (links to more, etc.) from captured searches.
- Run each captured result through Surfbot to retrieve actual page, making it possible to review quickly and consistently.

Note: The analyst considered the first fifty hits on each search for content relevancy. All initial searches were done on the same day; subsequent Surfbot retrievals took two days to complete.

Ultraseek Search Engine

The Ultraseek Search Engine has a search default of OR. If the searcher places more than one word or a phrase in the search box, the results will rate pages that include all words highest in relevance, and pages with at least one of the search terms lowest. If the searcher uses double quotes around a phrase, the results will list only pages that include all terms quoted. The search engine spiders many types of documents beyond the typical html, pdf, asp, and cfm extensions. Ultraseek allows tuning of search results by weighting certain html metadata tags. This includes the leveraging of Dublin Core elements such as title, subject (keyword),

Findings and Recommendations

The number of state agencies with discovered Web sites increased greatly from the first pass of testing to the second, and somewhat from the second pass to the third. Further, a wider variety of sites within these agency's scope were found. In addition, the analyst found a "good mix" of agencies within a search topic. This is important, revealing the prevalence and accuracy of the Dublin Core metadata application as well as increasing the possibility of searchers gleaning a broad range of information on a topic. On the negative side, however, a search term that was part of an agency name; e.g., "environment", caused results to be skewed in the direction of the particular agency. Generally, however, total number of hits as well as those of quality is increased. A summary of spreadsheet results for the three passes can be viewed at: http://bridges.state.mn.us/user2spreadsheet.pdf.

Use of quotes by the analyst to search for phases appears to be quite successful. Often, a higher number of hits from first to second to third passes were a result of using this method, as well as a larger number of relevant hits on the exact search phrase as well. This indicates greater search discovery due to the inclusion of Dublin Core in a Web site's source code. Single-word terms also resulted in better search results. For example, second and third passes searches using terms such as "agriculture," "soils," and "subsidies" had a greater number of total hits.

The "Relevance Ranking" field of the spreadsheet, utilized in passes 2 and 3, contains interesting results as well. This information, automatically generated by Ultraseek, depicts the quality of the site found in relation to the terms with which it was searched. The analyst recorded scores of the first hit as well as the fiftieth, with the former usually being far higher in number than the first. Comparison of results again indicates that the added Dublin Core increases search accuracy.

On another positive note, the analyst received very few "file not found" messages. In fact, less than 50% of 100 searches had any bad results. Further, on the second pass of testing, only a single term has no results whatsoever ("Anura," which has been eliminated from the report spreadsheet). This would indicate that Minnesota state agency Web sites are being frequently and consistently spidered by the Ultraseek, providing current and relevant information.

Analyst's Comments

Concluding the analyst's report after her second and third passes of testing were various comments, observations, ideas and recommendations. These add insight to the work already accomplished by the Foundations Project, as well as point out aspects that could still be improved. Included are:

Quality of Search Results:

1. By the third pass, the top 50 items found by both quoted and unquoted searches are quite similar. Also in the final pass, the best matches appear on the first page of results regardless of the search type or terms.

3. In the second and third passes, the mix of agencies covering the search topic is quite good, especially for policy-related items.

4. The top and bottom "relevance ranking" scores for each set of search results (second and third passes) continues to reflect the actual quality of overall results fairly well.

5. There is less of a problem with inappropriate pdf files being found on the second pass and especially on the third pass. Those discovered are more applicable to the terms searched than with initial searching. Some agencies, the analyst notes, have maintained the habit of dumping their files into pdf from the first pass to the second. Nonetheless, these files covered policy issues, and were in a format that didn't need a plug-in. By the time the third pass was conducted, Foundations staff had begun the use of pointer files, small html files that holds the metadata for pdfs, and links to them.

6. Another comment indicating the existence of very large files in search results, from both the first and second testing passes.

Agency Comments and Recommendations:

1. State agencies should be incorporating the metatag index directive more effectively.

2. A comment about slow links and downloads from some state agency servers.

3. A number of irrelevant files were indexed. They include: site logs and feedback forms, quasi-resumes, video catalogs and press releases. Since the agencies are responsible for what gets indexed, this is a matter controlled by each specific Web site. However, by pass 3, few inappropriate file types (site logs and comment forms, e.g.) were found.

4. Several raw directories were found in both first and second pass searches. These are documents without an index file or an explanation of their purpose, and thus not helpful to searchers. Again, by pass 3, no raw directories were found.

Discussion

After the third pass, it is clear that the major task of the Foundations Project has been successful; cataloging of Minnesota state agency Web sites with Dublin Core metalanguage has resulted in improved search results. Further, with higher numbers of Web sites cataloged and spidered, development of the LIV-MN thesaurus, and Ultraseek search engine's ability to use quotes and Boolean symbols, search results have improved in both quantity and accuracy. Consider also that the analyst's searches were done very simply, using either terms from the LIV-MN or keywords. With good results from such basic and replicable searching, it becomes evident that both experienced and non-expert, non-trained users would have little problem achieving abundant and accurate search results. Important, too, is the broad spectrum of state agencies that had information on searched topics.

Cost-Benefit Findings

Relevant to the above analysis, is the cost of adding Dublin Core metadata to web pages. The benefits of adding dc.subject tags, utilizing a controlled vocabulary, are clear. From internal studies, the Foundations Project staff have found it takes an average 5-7 minutes per page to add metadata. The lower time is due to familiarity with the subject matter, for example the page creator who already has worked in the subject field. The greater time reflects pages that are more difficult to analyze. In this range, it is assumed the person applying metadata has established familiarity with TagGen, the metatag software, and LIV-MN, the controlled vocabulary. Among the eight Foundations staff metadata catalogers, the average pay rate is $17/hour. Assuming a 75% production rate, the cost comes out to between $1.12 and $1.50 per web document. A note about granularity: not all documents on a Web site need to or should have metadata. The goal of metadata is to get users to the most relevant documents on a specific topic. Often that topic is found on an index.html page, with links to allow further exploration of that topic. In this case, the metadata belongs on the index page, and the user can walk the rest of the way down if more specificity is desired. Another example is periodicals that are mounted on the Web. Here the goal is to get the user to the front page of the periodical, which may be a template containing the latest issue, or an opening page that has the archives listed. Then, as each new issue is added, the metadata remains constant on the opening page or the template page, and needn't be added to each subsequent issue. Because of these rules of thumb, Bridges has added metadata to approximately the top 50% of the Web pages spidered on the Bridges search site. This makes the addition of metadata much more cost-effective than if each page had to have metadata.

Bibliography

Bachiochi, D. et al. "Usability Studies and Designing Navigational Aids for the World Wide Web." (June 2000).
http://www.rpi.edu/~danchm/Pubs/Web%20Paper/PAPER1~2.HTM

Buckingham Shum, S. "The Missing Link: Hypermedia Usability Research & the Web." *Interfaces, British HCI Group Magazine*, Summer 1996.

Goldsborough, Reid. "Searching the Web." *RN*, 62i3, March 1999, 23(2).

Goldwasser, Romi & Izarek, Stephanie. "Structural Strategies." *Computer Shopper* 16:16, December 1997, 626(3).
http://computershopper.zdnet.com/

Hawking, David et al. "Results and Challenges in Web Search Evaluation." (September 27, 1999).
http://www8.org/w8-papers/2c-search-discover/results/results.html

Henzinger, Monika et al. "Measuring Index Quality Using Random Walks on the Web." *Proceedings of the 8th International World Wide Web Conference*, Toronto, Canada, pp. 213-225. Compaq Computer Corporation, 1999.
http://www8.org/w8-papers/2c-search-discover/measuring/measuring.html

Instone, Keith. "How to Test Usability." (December 9, 1999).
http://instone.org/howtotest/index.html

Instone, Keith. "User Test Your Web Site: An Introduction to Usability Testing." (March 10, 1999). Originally appeared in *Web Review*, April 1997.
http://instone.org/howtotest/introduction.html

Jenkins, Charlotte et al. "Automatic RDF Metadata Generation for Resource Discovery." (June 9, 1999).
http://www8.org/w8-papers/2c-search-discover/automatic/automatic.html

McIntyre, Micki. "Finding a Syringe in a Haystack: Analyzing Search Engine Results." *Internet Librarian '99, Proceedings*. Medford, NJ: Information Today, 1999, pp. 107-109.

Navarro-Boomsliter, M. L. "Usability Testing--Expectations Versus Reality." *STC Proceedings*, 1992, p. 83.

Navarro-Prieto, Raquel, Scaife, Mike, & Rogers, Yvonne. "Cognitive Strategies in Web Searching." Human Factors & the Web [conference]. (June 2000).

http://zing.ncsl.nist.gov/hfweb/proceedings/navarro-prieto/

Pejtersen, Annelise Mark & Fidel, Raya. "Framework for Cognitive Work Analysis." Working Paper for MIRA Workshop, Grenoble, March 1998.
http://www.dcs.gla.ac.uk/mira/workshops/grenoble/fp.pdf

Rubin, Jeffrey. *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests.* New York: John Wiley & Sons, 1994.

Saracevic, Tefko et al. "A Study of Information Seeking and Retrieving." *Journal of the American Society for Information Science*, 39(3), 1988, pp. 161-176.

Shneiderman, B., Nyrd, D., & Croft, B. "Clarifying Search: A User Interface Framework for Text Searches." *Dlib Magazine*, January 1997.

User Interface Engineering. "Observing What Didn't Happen." (March 23, 2000). Originally appeared in *Eye for Design*, January/February 1999.
http://world.std.com/~uieweb/observng.htm