

Semester Project Final Report

Elisabeth Riba
May 5, 2003

Table of contents

Description of the collection	1
Range of coverage	1
User population	2
The need for subject access	2
Examination of other indices - similar format	3
Index format	5
Examination of other indices - similar content	6
Procedures, problems and progress	7
Evaluation	10
Attachments	10
Bibliography	10

Note: Text of this report available online at <http://www.osmond-riba.org/lis/419/Final.pdf>.
Links are not live, but may be copied onto the clipboard and pasted into your browser.

Description of the collection

A weblog, or blog is "a web page that contains brief, discrete hunks of information called posts. These posts are arranged in reverse-chronological order (the most recent posts come first). Each post is uniquely identified by an anchor tag, and it is marked with a permanent link that can be referred to by others who wish to link to it." (Doctrow et. al., 2002)

Joshua Micah Marshall is a writer living near Washington, DC who recently earned his doctorate in history. In November, 2000, he began writing a political blog, *Talking Points Memo* (abbreviated as *TPM*) as a forum to comment upon the contested presidential election. After the election was decided, Marshall continued blogging, writing primarily about national politics and foreign affairs.

Marshall once described his site as follows:

Talking Points Memo concerns itself with American politics, grand strategy abroad, Enron, snarky comments about the likes of Maureen Dowd and Howard Fineman and the admittedly rather obscure matter of whether a gentleman named Douglas H. Paal will ever be appointed -- as long rumored -- to serve as the Director of the American Institute in Taiwan, America's de facto embassy in Taiwan.

(April 4, 2001. <http://talkingpointsmemo.com/april0201.html#0404021043am>)

Talking Points Memo was one of the earliest political news blogs and has become quite influential. It's quoted regularly in the mainstream press and has been credited with breaking the recent Trent Lott scandal.

Range of coverage

The entire two-and-a-half year archive proved to be too large for such a project. After considering several methods of segmenting the collection, this author chose to index six months' worth of entries, from September 11, 2001 through September 11, 2002. Although this range spans the middle of the collection, this passage offered several advantages over other possible divisions:

- Focusing on the most recent entries is impossible because the weblog is ongoing, presenting the indexer with a constantly moving target.
- Starting from the beginning offered a different set of difficulties. For its first four months, *Talking Points Memo* did not provide anchor links to individual posts. Considering that some web pages have over twenty posts, anchor links are essential to a usable weblog index. In addition, this author thought these early pages also felt unfocused, and somewhat dated. Many external links pointed to sites that no longer exist. Aside from certain major events, there may not be much demand for an index of these entries.
- Other subdivisions appeared either too arbitrary or resulted in a collection that was still too large to be indexed.

Looking at posts from the milestone to its six month anniversary provides an interesting thematic focus to the collection. This time span also holds a certain historical significance, making an index of its contents more relevant to students, historians, political scientists, and other researchers.

The collection is still significant. The 181 days covered consist of 25 web pages with over 300 entries. Even when all content except the entries are eliminated (removing headers, advertisements, the side bar, Mr. Marshall's bio and links to the archive) and the formatting is adjusted so entries span the full length and width of the page, it still contains nearly 200 pages of text.

For now, this is intended as a finite collection examining a specific historical range. Consideration of whether to expand the coverage further will be discussed with Mr. Marshall.

User population

Potential users for such an index fall into two groups: regular readers of *Talking Points Memo* and researchers studying the political landscape in the aftermath of September 11, 2001.

Examining the user population of a weblog is extremely difficult. A profile of the *TPM* readership can only be derived from Marshall's comments about his readers and from other people who link to or quote him. Overall, it appears that *TPM*'s audience is primarily college-educated Americans actively interested in politics, generally coming from a liberal perspective. A recent profile of Marshall supports this analysis ("Here's what I think," 2003): "Marshall is well acquainted with his readers: He gets 100 to 150 emails a day from them. From anecdotal evidence, Marshall's readers appear to be political junkies who mostly live outside Washington — 'a mix of professionals, lawyers, and academics,' he says."

Researchers interested in this collection would include journalists, political operatives and academics. This author assumes most such users would have at least some college education, although it's possible that a few high school students might find this a useful resource.

The need for subject access

Talking Points Memo currently has no subject index, although Marshall does provide users with a broad date index and a search engine.

The bottom of each web page in the blog lists and links to all the pages in the archive. (See figure 1). Searchers who know which week an entry was posted can thus go directly to that page, although that still leaves the issue of finding a particular entry on the page.

TPM does include search capabilities powered by Atomz (<http://talkingpointsmemo.com/advancedsearch.html>, figure 2). Although it has some useful features, including sound-alike matching, search results only link to the web page, again leaving it up to the users to find the specific entry.

This author also uses Atomz to index her weblog, so is familiar with its capabilities. The index is updated once a week at most, making it difficult to find the most current information. Atomz does provide site owners with regular reports listing all search terms users have entered. Although this would be a useful resource to evaluate the users' vocabulary, requests to Mr. Marshall for copies of these reports have gone unanswered.

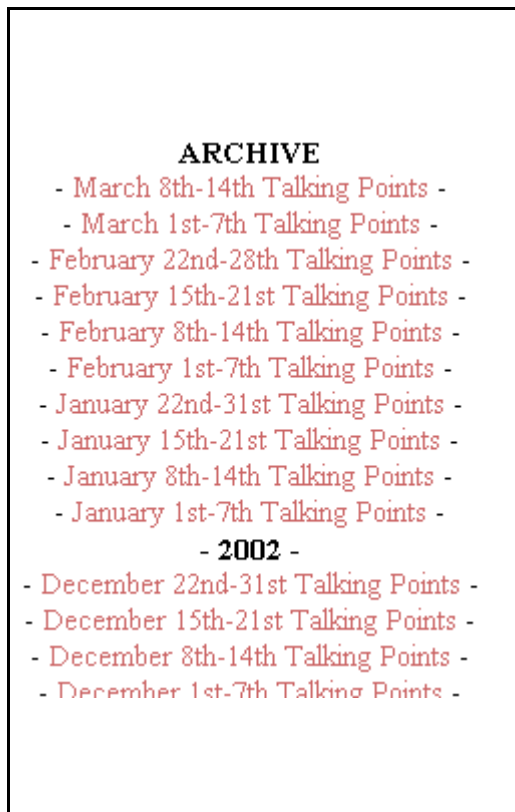


figure 1: TPM archive links

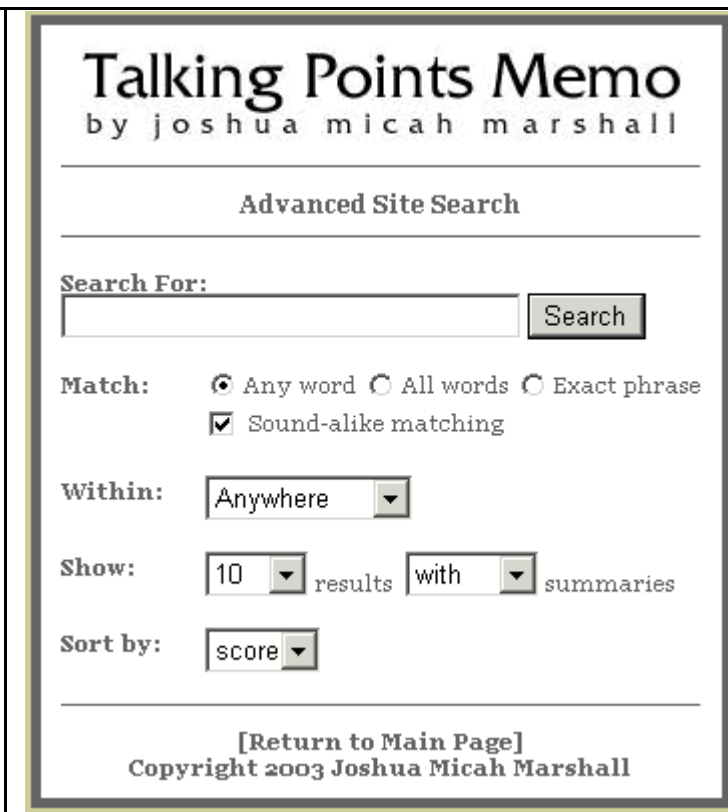


figure 2: TPM search page

Users can also use other commercial search engines to find entries within the website. Again, this is limited by the frequency of indexing and will only take users to the page, not to a particular entry.

Also, these search services rely strictly upon keyword searching. Atomz offers sound-alike matching, but sometimes the desired terms do not appear in a particular post. A recent post about the Columbia disaster (outside the scope of this project) provides a good example:

AN IMPORTANT POINT for me to add regarding the Easterbrook **piece** mentioned below. Definitely read it. It's important. But I don't share the lack of enthusiasm about manned space flight which comes through in his new **piece** today in Time.com. Nor do I agree with all the particulars in the exhaustive earlier *Monthly* article. Just wanted to make that clear. More soon.

(February 2, 2003. <http://talkingpointsmemo.com/feb0301.html#020203432pm>)

This is the entire entry. Notice that Marshall never uses the words "Columbia," "NASA," or "shuttle." They do appear in subsequent posts on the same web page, so people searching on those words would find this page, but unless they pay close attention, it's easy to overlook this "important point."

Examination of other indices - similar format

Searching the Internet for information on weblog indexing uncovers many directories of weblogs, grouped by broad subject matter, political leanings, or the blogger's geographical location. Very little could be found on subject indexing within a single blog. Generally such indexes are created and maintained by the author of that blog.

Fury.com, Kevin Fox's hand-coded weblog, has a sidebar on the right of every page labelled "Read by Topic" (figure 3). Clicking one of these links reformats the page to display all entries associated with that category (figure 4).



figure 3: Fury.com sidebar

figure 4: Fury.com: selected topic

Movable Type (MT), a popular software package used to generate blogs, includes categorization as one of its features. As with *Fury.com*, bloggers can assign categories to their posts. For archives by category, MT merely lists the matching posts, sorted by date and using the title field as links (figure 5).



figure 5: DailyKos archive by category

Looking beyond blogs, many websites do have site indexes. Unfortunately, examining them hasn't proven terribly useful. In almost every case, every index term corresponds to exactly one entry. Terms may be alphabetical (figure 6) or hierarchical (figure 7), but this one-to-one correspondence will not work for a weblog, in which one topic may be discussed over a period of weeks or months. Only one web index was found making multiple links from a single term -- the W3C HTML 4.01 specification (figure 8). However, the links were completely nondescriptive numbering.

<p>Forms (OSHA Publications) Freedom Of Information Act Frequently Asked Questions Frequently Cited OSHA Regulations</p> <p>G TOP General Industry Standards and Regulations General Safety & Health General Safety & Health GOCAD - Cadmium Standards Government Employees Grain Handling Grants, Susan Harwood Grocery Warehousing</p> <p>H TOP Hand and Power Tools Hazard Awareness and Communication Hazard Communication Hazard Information Bulletin Hazardous Drugs</p>	<p>News</p> <p>Nation Columns/Caricatures Courts DOT.MIL National Security Science Search the States Special Reports Photo Galleries Live Online Nation Index</p> <p>World Search the World Special Reports Africa Americas Asia/Pacific Europe Middle East Live Online</p>	<p>Travel Escapes Visitor's Guide What to See Where to Go Dining Getting Around Live Online Online Travel Heritage Columns Travel In</p> <p>Health Alternative Children's Health Chronic Diseases Fitness Health Care Men's Health Mental Health</p>	<p>boolean attribute 1, 2 minimized 1</p> <p>border around a frame 1 around a table 1 around image 1 around object 1</p> <p>cascading style sheets 1</p> <p>case of URIs 1 of anchor name 1 of attribute names 1 of attribute values 1, 2, 3 of character encodings 1 of character entity references 1</p>
--	--	--	---

figure 6: alphabetical site index, OSHA.gov

figure 7: hierarchical site index, Washingtonpost.com

figure 8: index terms with multiple links, W3C.org

Most articles on how to index websites echoed these deficiencies, and were irrelevant to the task at hand.

Index format

To be truly useful, a weblog index must somehow connect users from the terms to the respective entries. Thus, the final product will be written in HTML with links from the index terms to the individual posts.

This will not be fully integrated like that of *Fury.com*. Such a task would require major rewrites to the existing site. This author hasn't the access rights, coding skills, or time in which to undertake such a project. What's more, such extensive revisions have the potential to break links or otherwise damage existing functionality.

This index will be a separate web page and must be able to operate independently of the *Talking Points Memo*. It will probably be hosted on this author's domain (<http://www.osmond-riba.org/>), although the completed product will be offered to Mr. Marshall for inclusion on his website.

A format similar to that generated by Movable Type might be feasible. However, that layout has several problems both as applied to the content of *TPM* and as related to desirable index properties.

Unlike posts generated by MT, Marshall's blog entries do not have titles. He does distinguish the first three words of each post with a different HTML style. However, these headings are useless from an indexing point of view because they are neither unique nor descriptive ("This is a" is used six times).

Also, most MT indices use a small number of broad categories. The index shown in Figure 5 assigned over 150 entries to the category "Bush administration." The index planned for *TPM* will be much more specific. Following the MT format would result in a huge number of web pages, most with only a handful of links. The need to go back and

forth between the term list and links hinders browsability and might frustrate users. This format also hides some of the information users need to differentiate between various terms (Maislin, 2003).

A better approach would be something closer to a book index, which lists both the terms and references on the same page. The page will include an alphabetical list of the index terms, with links from each term to the related entries.

Links to records will be represented in the index by their date and time. In blogs, date and time hold a similar significance to page numbers in printed material. Marshall's method of including the full date and time -- for example, "[September 11th, 2002 -- 4:03 PM EDT](#)" -- is too long and busy for an index which will probably list many such links in succession. Other attempted date/time formats also looked too busy. [It would be nice to list each link in a two-line format, with date displayed *over* time. However, that would require use of tables or images for each entry, which have a host of complexities of their own.]

References only list the dates, *except* when two adjacent links refer to the same date. In those cases, time is displayed to separate those entries. This means that the same entry is sometimes be listed by date and time and in other places listed only by date (See figure 9). This may make it difficult for users to cross-reference entries, such as recognizing which entry about Pakistan is also about India, however test users saw no problems with the format.



figure 9: date format in index entries

Seth Maislin has written in several places about the importance of page numbering in indices. (Maislin, 2003) (Maislin, 1999) (Brenner, 2000) Users benefit by seeing references in context. Page numbering can help users identify the more important mentions and recognize outliers or later updates. The date-time stamps in *TPM* can serve a similar purpose. Because sometimes adjacent posts may appear on different pages, a simple page range will not work.

Efforts have been made to resemble *TPM's* current user interface, but fidelity was sacrificed for usability. For example, Marshall restricts his content to a 300-pixel wide column. While this looks very nice for his text, the index makes more efficient use of the screen real estate.

Examination of other indices - similar content

Newspapers have historically been called "the first draft of history." Blogs fill a similar niche, even rougher than newspapers. Although other web indices proved incompatible for comparison purposes, an examination of newspaper library practices was more fruitful.

Most books on news librarianship focused more on the management side than on indexing practices. However, the author was able to find several relevant chapters for reference during the indexing process.

This author also examined some existing news indices, including the *Readers' guide to periodical literature*, and several by UMI (which indexes *New York Times*, *Boston Globe*, and numerous other major dailies). Although moderately interesting, these were not terribly useful before the author began the indexing process.

Procedures, problems and progress

The early stages of the project were largely technical ones. The author mirrored the entire contents of the <http://www.talkingpointsmemo.com> website onto her personal computer, using BlackWidow (available from <http://www.softbytelabs.com/>). This eliminated network connectivity risks and enabled the author to view and alter the HTML code for these pages.

Talking Points Memo was designed to be read online in small chunks. The formatting was not terribly conducive for lengthy printouts or sustained reading. For example, the blog entries are written in a column 300 pixels wide. In letter paper, the margins are wider than the text, both a huge potential waste of paper and difficult to read. Similarly, each page concludes with a date index and Mr. Marshall's biography. Those would consume four sheets of letter paper -- for each web page.

The author manipulated the HTML code (though didn't touch Marshall's text) to provide maximum space for the actual blog entries and eliminate everything else. These edits reduced the page count by nearly fifty percent, as concatenated via Adobe Acrobat.

A spreadsheet program (Lotus 1-2-3) was used to extract the bookmark links, dates and first three words from all the posts in the range. These were used to generate some tables of contents for the collection, both in list format and a calendar view. This process also uncovered a few duplicate entries, which the author was able to remove from the collection.

The indexer then read through these pages, writing down concepts she thought appropriate for each entry. Rather than using index cards, this was done into a relational database (Lotus Approach) consisting of two tables: one derived from the spreadsheet described above, with concepts in a many-to-one relationship to that first table (See figure 10).

Num	First3Words	<input type="checkbox"/> 1st Batch	DTString	URL
155	John Walker's parents		December 20th, 2001 -- 5:41 AM EST	http://talkingpointsmemo.com/dec0103.html#122001541am
Heading		Subhead		Notes
Lindh, John Walker				Terms
military tribunals				3
civil liberties		right to a lawyer		

figure 10: joined databases used to generate concepts

Once all entries had been assigned concepts, the records were sorted by the concept field, and new fields were added for the index term. Concepts were refined into index terms, consolidating some and trying to make them consistent (See figure 11).

200	January 14th, 2002 -- 12:40 AM EST	Gramm, Phil	Enron dealings; why he retired	Gramm, Phil
114	November 28th, 2001 -- 10:02 AM EST	Greece		Greece
116	November 28th, 2001 -- 7:51 PM EST	Greece		Greece
120	November 29th, 2001 -- 8:44 PM EST	Greece		Greece
81	November 5th, 2001 -- 9:13 AM EST	Gurung, Subash		Gurung, Subash
82	November 6th, 2001 -- 11:29 AM EST	Gurung, Subash		Gurung, Subash
83	November 6th, 2001 -- 5:15 PM EST	Gurung, Subash		Gurung, Subash
311	March 8th, 2002 -- 11:18 AM EST	Helms, Jesse		Helms, Jesse
12	September 19th, 2001 -- 8:02 PM EST	Hussein, Saddam		Hussein, Saddam
128	December 4th, 2001 -- 1:26 AM EST	Hussein, Saddam	toppling/overthrowing	Hussein, Saddam
146	December 13th, 2001 -- 11:34 AM EST	India		India
168	December 27th, 2001 -- 5:47 PM EST	India		India
180	January 3rd, 2002 -- 7:40 PM EST	India	relations w/Pakistan	India

figure 11: concepts used to generate index terms

Whereever possible, the author tried to stick close to Mr. Marshall's terminology. Whatmore (1978) was an invaluable resource in formulating appropriate subject headings, as was UMI's controlled vocabulary (Bell & Howell, 2000) . These were used merely for guidance, and neither was followed strictly.

For the interim report, the author indexed approximately ten percent of the entries. These were chosen in a non-random method, using the spreadsheet to select every tenth entry, including the first and last in the date range. In addition, several adjacent entries were selected for design purposes (how to distinguish such entries).

There were problems with this methodology. For one thing, stories often developed over time. Sometimes a subject would receive heavy coverage for several posts in a row and then vanish. Some subjects that seemed irrelevant when a post was viewed in isolation came into prominence when viewed in context. After further study, it was realized that the sample index of every tenth record did not give an accurate image of those entries. Although the interim index was used as an occasional reference, the final version reindexed all the entries in hopes of better consistency overall.

[Another complication between the interim and final reports involved the interference of real life events into the assignment. Personal concerns over finances and jobhunting became major distractions. Although these worries were eased when the author finally obtained full-time employment, the time commitment of the new job severely limited the time available to complete this project. As a result, the final index coverage was reduced from one year to six months.]

For the final report, the author also used the thesaurus feature of Inmagic DB/TextWorks to manage RT and Use references (See figure 12). Experiments were done in migrating the entire database to DB/TextWorks, migrating the thesaurus format to Lotus Approach, and trying to make the two programs interoperate (through ODBC). In the end, they were run as separate applications.

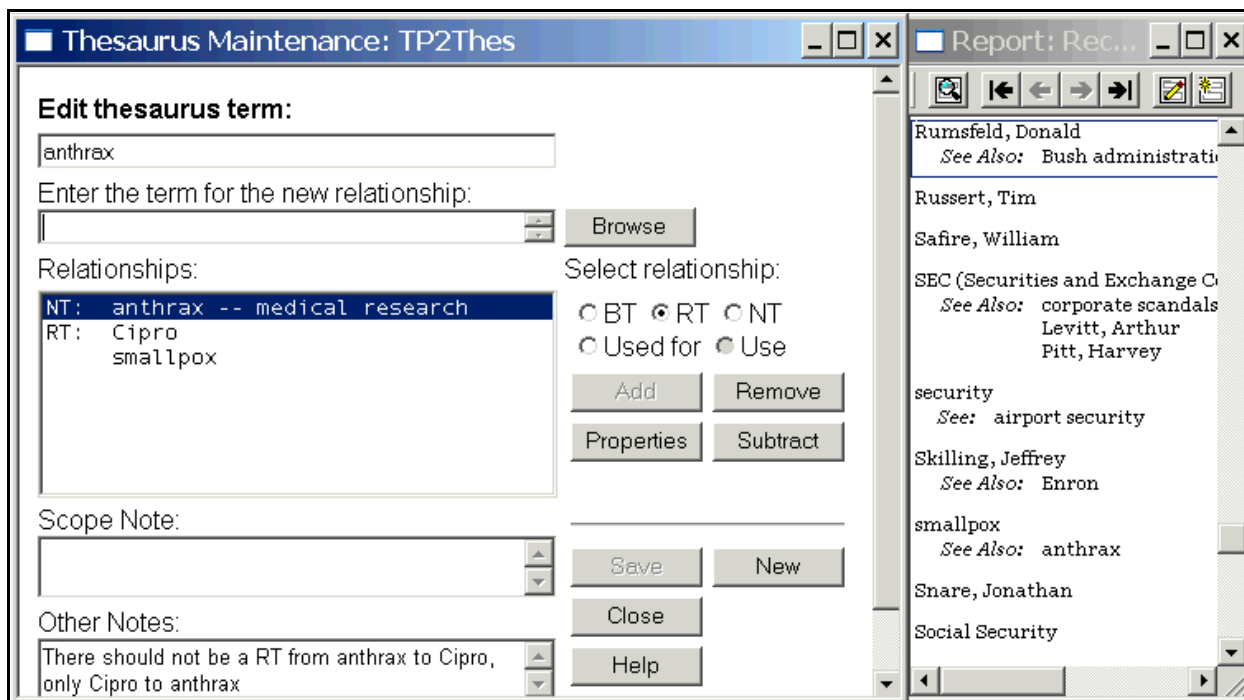


figure 12: DB/TextWorks thesaurus of index terms

After sufficient terms were complete, the resulting tables were brought back into the spreadsheet, where formulas were used to generate the HTML for the body of the index (See figure 13) which was then pasted into a text editor and saved as HTML. The author codified the HTML format after approximately one-third of the terms were complete to verify the appearance and usability of the index.

See	See AlstNotes	Term	Date	Time	URL	dstr	tstr	dcount	alpha?	DT	DD
	bin Laden, Osama	Abu Ghaith, Suleiman						3	</DL>*alph	<DT><A NA	<DD> See also <A
		Abu Ghaith	12/13/01	11:54 PM	http://talkin	12/13/01	11:54 PM	3			 <A CLASS="ref" HREF="http://talk
		Abu Ghaith	12/27/01	07:12 PM	http://talkin	12/27/01	17:12 PM	3			: <A CLASS="ref" HREF="http://talkin
	Karzai, Hamid	Afghanistan						3		<DT><A NA	<DD> See also <A
	Northern Alliance	Afghanistan						3			: <A HREF="#north
	Taliban	Afghanistan						3			: <A HREF="#talib
	bin Laden, Osama	Al Qaeda						1		<DT><A NA	 See also <A HREF
	plagiarism	Ambrose, Stephen						5		<DT><A NA	<DD> See also <A
		Ambrose, S	01/08/02	08:37 AM	http://talkin	1/8/02	08:37 AM	5			 <A CLASS="ref" HREF="http://talk
		Ambrose, S	01/08/02	12:00 AM	http://talkin	1/8/02	12:00 AM	5			: <A CLASS="ref" HREF="http://talkin
		Ambrose, S	01/09/02	11:25 PM	http://talkin	1/9/02	11:25 PM	5			: <A CLASS="ref" HREF="http://talkin
		Ambrose, S	01/10/02	08:37 PM	http://talkin	1/10/02	08:37 PM	5			: <A CLASS="ref" HREF="http://talkin
	smallpox	anthrax						12		<DT><A NA	<DD> See also <A
		anthrax	10/17/01	02:56 PM	http://talkin	10/17/01	02:56 PM	12			 <A CLASS="ref" HREF="http://talk
		anthrax	10/19/01	01:51 PM	http://talkin	10/19/01	01:51 PM	12			: <A CLASS="ref" HREF="http://talkin
		anthrax	10/23/01	03:19 PM	http://talkin	10/23/01	03:19 PM	12			: <A CLASS="ref" HREF="http://talkin
		anthrax	10/24/01	01:13 AM	http://talkin	10/24/01	01:13 AM	12			: <A CLASS="ref" HREF="http://talkin

figure 13: spreadsheet used to generate the HTML

The index was sorted in alphabetical order, with subheadings for some of the more frequent terms. Personal names were indexed by surname. This format keeps President Bush sorted near his brother and articles about his administration.

Evaluation

A matrix of entries and concepts were generated, and the resulting arrays were examined for duplication.

Several test subjects fitting a similar profile to the target audience -- college educated and politically aware -- were recruited for usability testing. These selected users were given the URL to the index and asked for their opinions on the suitability of terms and the index format.

Compliance with HTML and accessibility standards was verified using <http://validator.w3.org/>. To ensure the UI would display adequately, the author also viewed the page using four popular browsers: Opera, Internet Explorer and Netscape Navigator under a Windows client, and Lynx (non-graphical) from a UNIX client. The offsite usability testers were running on an Apple iMac.

Attachments

- The final index may be viewed online at <http://www.osmond-riba.org/lis/419/TPM.htm>

Bibliography

- Bear, N. Z. (2003). *Blogosphere Ecosystem*. Retrieved from <http://www.truthlaidbear.com/ecosystem.php>
This is a completely automated application which scans weblogs once daily and generates a list of weblogs ranked by the number of incoming links they receive from other weblogs on the list. All links from a scanned weblog to any other weblog (i.e., links from a weblog to itself are ignored). This includes links within posts, and 'permanent' links in a weblog's blogroll. Although its list is far from exhaustive, it does provide a nicely sized sample. And, unlike other popular blog indexes, this does not depend upon RSS feeds (which *TPM* lacks).
- Bell & Howell Information and Learning. (2000). *Controlled vocabulary of subject terms*. Retrieved from <http://www.umi.com/hp/Support/Vocab/>
- Borko, H., & Bernier, C. (1978). *Indexing concepts and methods*. New York : Academic Press
- Brenner, D., & Rowland, M. (2000). *Beyond book indexing: How to get started in Web indexing, embedded indexing, and other computer-based media*. Medford, NJ : Information Today, in association with the American Society of Indexers.
Chapters 3 - 6 were most relevant to this project.
- *Daily Kos: political analysis and other daily rants on the state of the nation*. (2002-2003). Retrieved from <http://www.dailykos.com/>
This is a political blog with a subject index.
- Doctrow, C., Dornfest, R., Johnson, J. S., Powers, S., Trott, B., & Trott, M.G. (2002). *Essential blogging*. [Electronic version]. Sebastopol, CA : O'Reilly. Retrieved from <http://safari.oreilly.com/?XmlId=0-596-00388-9>
This book was only used for its definition of a weblog. The authors and publishers have a good reputation for technical accuracy.
- Drum, K. (2003, February 7). [Interview with Joshua Micah Marshall] *CalPundit*. Retrieved from http://calpundit.blogspot.com/2003_02_02_calpundit_archive.html#88695099
- Fox, K. (2000-2003). *Fury.com*. Retrieved from <http://www.fury.com>
This is a blog with a superb subject index including unique implementation.

- Here's what I think. (April 23, 2003). *Princeton Alumni Weekly*. [Electronic version]. Retrieved from http://www.princeton.edu/~paw/web_exclusives/alumni_spotlight/as_042303marshall.html
Appearing just before my deadline, this profile of Mr. Marshall confirmed my suspicions about his website's readership.
- International Organization for Standardisation. (1985). *Documentation -- Methods for examining documents, determining their subjects, and selecting index terms*. ISO 5963-1985(E).
- Lathrop, L. (2000, May). Index usability test questions. [Electronic version]. *A to Z*. Retrieved from <http://www.stcsig.org/idx/articles/usability.pdf>
- Maislin, S. (1999, summer). Indexing online: the new face of an old art. *Hyperviews Online*, 2(3). Retrieved from <http://taxonomist.tripod.com/websmarts/onlineindexing.html>
Page numbering provides context to users of print indices, but are often omitted from online indexes. This article demonstrates the importance of this information, and suggests alternative solutions to increase the utility of indexes online. The information in this article overlaps parts of *Beyond book indexing* (Brenner, 2000, chap. 5) written by Maislin.
- Maislin, S. (2000). Writing an index using Microsoft Excel. *Society for Technical Communication SIG newsletter*. Retrieved from <http://taxonomist.tripod.com/indexing/excelindexing.html>
This article provided techniques for indexers who wish to sort and export their output using a spreadsheet program. Since this author already has a tendency to use spreadsheet tables and formulas to generate HTML code for other purposes, it seemed like a good match on first glance. Unfortunately, this article was geared more towards novice users; this author is already quite familiar with the techniques Mr. Maislin described.
- Maislin, S. (2003). Frequently asked questions on indexing and information architecture. Retrieved from <http://taxonomist.tripod.com/faq.html>
The FAQ contains short snippets of information on a variety of topics, none of which were significant enough to fill an entire article. This has slightly updated information on page numbering and use of page ranges.
- Marshall, J.M. (2000-2003). *Talking Points Memo*. Retrieved from <http://www.talkingpointsmemo.com>
Note: In-text cites to individual posts will use a slightly different format, providing the date and anchor link, but omitting Mr. Marshall's name.
- Riba, E. (2000). Amazon.com vs. INNOPAC: which interface is easier to search? Retrieved from <http://www.osmond-riba.org/lis/proposal.htm>
Written by this author several years ago, this paper proposed a research project (never conducted) to compare the web-based user interfaces of Amazon.com bookseller and INNOPAC's library catalog. The paper cited several sources on the number of users needed for usability testing. Rather than redoing this research, the author has decided to refer back to her earlier work, which quoted the following articles:
 - Chisman, J., Diller, K., & Walbridge, S. (1999, November). Usability testing: a case study. *College & research libraries*, 60(6), 552-569.
 - Nielsen, J., & Landauer, T. (1993, April 24-29). A Mathematical model of the finding of usability problems. *Proceedings of ACM INTERCHI '93 conference*. Amsterdam : ACM. 206-213.
 - Nielsen, J. (2000, March 19). Why you only need to test with 5 users. Retrieved from <http://www.useit.com/alertbox/20000319.html>
- Rosenfeld, L. (1997, October 3). Organizing your site from a-z. *WebReview.com*. Retrieved from http://www.webreview.com/1997/10_03/strategists/10_03_97_4.shtml

- Schwartz, C. (2001). *Sorting out the Web: Approaches to subject access*. Westport, CT : Ablex.
- Semonche, B. P. (1993). Newspaper Indexing Policies and Procedures. [Electronic version]. In B. P. Semonche (Ed.), *News media libraries: a management handbook*. (pp. 373 - 412). Westport, CT : Greenwood Press. Retrieved from <http://www.ibiblio.org/journalism/indexing.html>

Although this author did scan the full book, only this one chapter was relevant to the project, and of the chapter, only a third or so dealt with how to index.
- Special Libraries Association Newspaper Division. (1983). *Guidelines for newspaper libraries*. Reston, VA : American Newspaper Publishers Association Foundation.

Most of this book was relatively irrelevant to this project. Only chapters 11 (subject heading control) and 13 (indexing) were remotely helpful, but even those were biased towards managerial rather than practical advice.
- Steps for reviewing an index. [Electronic version]. In (1996). *Stet again! : more tricks of the trade for publications people*. Alexandria, VA: EEI Press. Retrieved from http://www.eecomunications.com/press/stet2/tocstet_indexing.html
- Whatmore, G. (1978). *The modern news library: documentation of current affairs in newspaper and broadcasting libraries*. London : Library Association.

Despite its age and foreign origin, this was one of the most helpful books on news indexing. Chapters 4 through 8 were particularly useful on how to classify, including suggested subdivisions for names and locations. "Sequences for special subjects" was most relevant, with advice on categorizing political issues and elections.